scientific reports

Check for updates

OPEN

COVID-19 detection from exhaled breath

Nicolò Bellarmino[1,5]✉, Riccardo Cantoro[1,5], Michele Castelluzzo[2,5], Raffaele Correale[2,5], Giovanni Squillero[1,5], Giorgio Bozzini[3], Francesco Castelletti[4], Carla Ciricugno[2], Daniela Dalla Gasperina[4], Francesco Dentali[4], Giovanni Poggialini[4], Piergiorgio Salerno[4] & Stefano Taborelli[4]

The SARS-CoV-2 coronavirus emerged in 2019 causing a COVID-19 pandemic that resulted in 7 million deaths out of 770 million reported cases over the next 4 years. The global health emergency called for unprecedented efforts to monitor and reduce the rate of infection, pushing the study of new diagnostic methods. In this paper, we introduce a cheap, fast, and non-invasive COVID-19 detection system, which exploits only exhaled breath. Specifically, provided an air sample, the mass spectra in the 10–351 mass-to-charge range are measured using an original micro and nano-sampling device coupled with a high-precision spectrometer; then, the raw spectra are processed by custom software algorithms; the clean and augmented data are eventually classified using state-of-the-art machine-learning algorithms. An uncontrolled clinical trial was conducted between 2021 and 2022 on 302 subjects who were concerned about being infected, either due to exhibiting symptoms or having recently recovered from illness. Despite the simplicity of use, our system showed a performance comparable to the traditional polymerase-chain-reaction and antigen testing in identifying cases of COVID-19 (that is, 95% accuracy, 94% recall, 96% specificity, and 92% F_1-score). In light of these outcomes, we think that the proposed system holds the potential for substantial contributions to routine screenings and expedited responses during future epidemics, as it yields results comparable to state-of-the-art methods, providing them in a more rapid and less invasive manner.

The COVID-19 pandemic, that began in late 2019, had an unprecedented global impact, with the World Health Organization (WHO) reporting over 770 million infections and more than 7 million deaths worldwide. The rapid spread of the virus prompted a global health emergency that lasted from January 2020 to May 2023, during which extraordinary efforts were made to monitor and reduce the infection rate. These efforts included widespread social restrictions, mass testing, and the development of vaccines and treatments to manage the crisis effectively[1,2].

Among the diagnostic tools utilized, real-time quantitative polymerase chain reaction (RT-qPCR) has been the gold standard for detecting SARS-CoV-2, the virus responsible for COVID-19. This technique relies on identifying viral ribonucleic acid (RNA) in nasopharyngeal or oropharyngeal swab samples, allowing for accurate and timely diagnosis[3,4]. Despite its widespread use, RT-qPCR has several limitations that can hinder the effectiveness of large-scale testing programs. The high sensitivity of RT-qPCR requires meticulous experimental design and a deep understanding of normalization procedures to avoid false-negative results, which can occur due to technical issues during sample collection, transportation, and processing, as well as biological factors like genetic variations, sample types, viral load, and the timing of sample collection relative to viral exposure[5]. Additionally, the necessity for authorized laboratories with at least Biosafety Level 2 (BSL-2) certification can place significant strain on laboratory resources, potentially leading to delays in processing and reporting test results. These challenges are compounded by the high costs associated with the equipment and reagents needed for RT-qPCR[6].

In light of these challenges, there has been a growing interest in developing alternative diagnostic methods that are rapid, cost-effective, non-invasive, and capable of detecting infections at an early stage[7–11]. One promising approach is the analysis of exhaled breath, which contains respiratory droplets and a variety of small molecules produced through metabolic and catabolic processes. Breath analysis has already been explored for the detection of several diseases, including lung diseases[9,10], breast cancer, diabetes, and infectious conditions such as influenza. Expanding the application of breath analysis to detect COVID-19 presents several significant advantages over traditional methods[12,13].

[1]Politecnico di Torino, Torino, Italy. [2]NanoTech Analysis Srl, Torino, Italy. [3]Hospital ASST Lariana, Como, Italy. [4]Università degli Studi dell'Insubria, Varese, Italy. [5]These authors contributed equally: Nicolò Bellarmino, Riccardo Cantoro, Michele Castelluzzo, Raffaele Correale and Giovanni Squillero. ✉email: nicolo.bellarmino@polito.it

In parallel, artificial intelligence (AI) has emerged as a particularly promising area of research for enhancing COVID-19 detection. AI has the potential to improve the accuracy and efficiency of diagnostics by analyzing complex patterns in various data types, including medical imaging, genomic sequences, and physiological signals. Recent studies have demonstrated the efficacy of AI models in analyzing chest X-rays, computed tomography (CT) scans, and even voice and cough sounds to identify COVID-19 infections with high accuracy[14,15]. Furthermore, AI-driven analysis of exhaled breath is gaining traction as a non-invasive and rapid diagnostic method. Machine learning (ML) algorithms can detect volatile organic compounds (VOCs) in breath samples that are indicative of SARS-CoV-2 infection. These AI-based methods could revolutionize COVID-19 diagnostics by enabling real-time, on-site testing that is both cost-effective and accessible[9,10,16,17]. However, related work mainly rely on analyze breath samples for specific VOC patterns.

This paper presents a novel approach to COVID-19 detection by integrating AI-based analysis with breath sampling techniques. The primary objective of this work is to evaluate the effectiveness of AI in detecting SARS-CoV-2 from exhaled breath, leveraging ML algorithms to detect the positivity to COVID-19 without the need for identify unique VOC signatures associated with the virus, but relying only on breath fingerprint. Unlike traditional RT-qPCR, this method aims to provide a rapid, non-invasive, and portable diagnostic solution that can be used in various settings, including high-traffic areas like airports and public transportation hubs. The developed system achieved results comparable to other classical COVID-19 detection systems: 95% accuracy, 94% recall, 96% specificity, and an F_1-score of 92%.

This approach not only addresses the need for faster and more comfortable testing methods but also offers a scalable solution that could be deployed in resource-limited environments. This system can be extended to other infectious conditions and diseases.

Code for spectra analysis and the outcomes of the experiments have been released in a public GitHub repository[1].

The organization of the paper is as follows: the *Method* section describes the method and experimental setup, including breath sample collection, pre-processing and AI model development. Section *Experimental Evaluation* presents the experimental evaluation, focusing on patients sample collection and performance evaluation of the predictive models. Section *Results* summarize the main results of the proposed approach in terms of final accuracy of the developed models. Finally, the *Conclusions* concludes the paper by summarizing the key contributions, potential limitations, and future research directions.

Method

We propose a detection system that leverages mass spectrometry and AI to rapidly assess exhaled breath samples from patients and identify the presence of COVID-19. Recent studies have shown that COVID-19 patients exhibit distinct VOC profiles in their breath[7,18,19].

VOCs are a significant group of chemicals that can easily evaporate at room temperature. They are present in various products, including paints, cleaning agents, and building materials. The expelled breath from individuals contains several VOCs in addition to nitrogen, oxygen, carbon dioxide, and water vapor. Recent studies have identified specific VOCs as biomarkers for several respiratory diseases, including lung cancer, cystic fibrosis, asthma, chronic obstructive pulmonary disease (COPD), and COVID-19. Furthermore, variations in VOC profiles can help distinguish between smokers and non-smokers. Certain VOCs have also been associated with lung cancer. For example, elevated levels of specific VOCs in exhaled breath have been correlated with lung cancer diagnosis, suggesting their potential utility in early detection[20]. In cystic fibrosis, VOCs may indicate disease severity and exacerbations, providing a non-invasive monitoring tool[20]. However, analyzing VOC necessitates the use of specialized techniques and hardware for detecting and selecting specific VOC[7,9,10,16,21,22]. Electronic nose technology and other analytical methods have demonstrated high sensitivity and specificity in detecting these compounds, making VOC analysis a promising tool for rapid COVID-19 diagnosis[7,9,10,21].

The proposed approach completely eliminates the need for prior identification of specific VOCs, focusing instead on the direct analysis of the breath *fingerprint* through its mass spectrum. This method is straightforward, easy to implement, and aims to establish a correlation between a specific breath fingerprint and the presence of COVID-19 without explicitly defining or detecting individual VOCs. Breath samples can be conveniently stored in specialized containers, simplifying collection procedures that can be performed by non-specialized personnel in various locations.

Our system utilizes a proprietary nano-sampling device coupled with a high-precision mass spectrometer capable of performing efficient mass spectrum analysis within the 10–351 m/z range[23]; this analysis requires usually few seconds, and never more than few minutes. The raw data are processed by in-house developed python tools: first they are aligned to the baseline, then filtered to reduce measurement noise, and eventually a process of data augmentation enhances the robustness and diversity. We employed standard ML classifiers from a state-of-the-art data analysis library[24] to detect the presence of COVID-19. Notably, this system operates without the need for reagents, and it generates no hazardous waste, making it both efficient and environmentally friendly.

Breath samples collection

For each patient under test, ambient air was sampled to verify environmental parameters and ensure the stability of the instrument. Then, the subject's breath is collected into a sampling tedlar bag with a defined volume of 3 L, by having the subject blow through a straw directly into the bag. Each patient exhales into the bag until it is filled with approximately 3 L of air. This large volume is necessary to establish the stable sampling pressure required

[1] https://github.com/BellaNico4/COVID-19-Detection-from-Exhaled-Breath

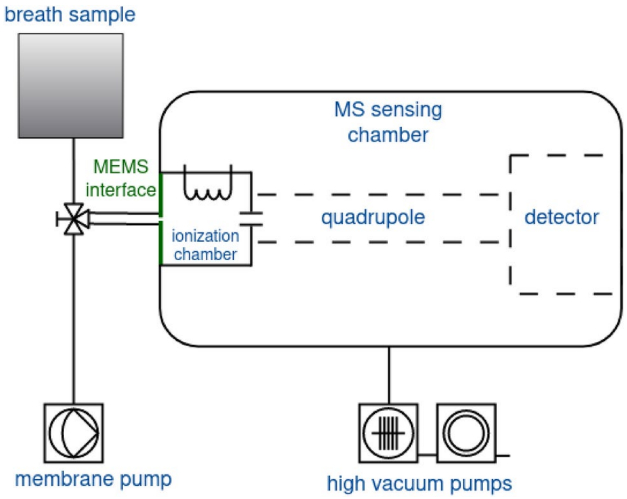

Figure 1. Schematic of the Mass Spectra analyzer.

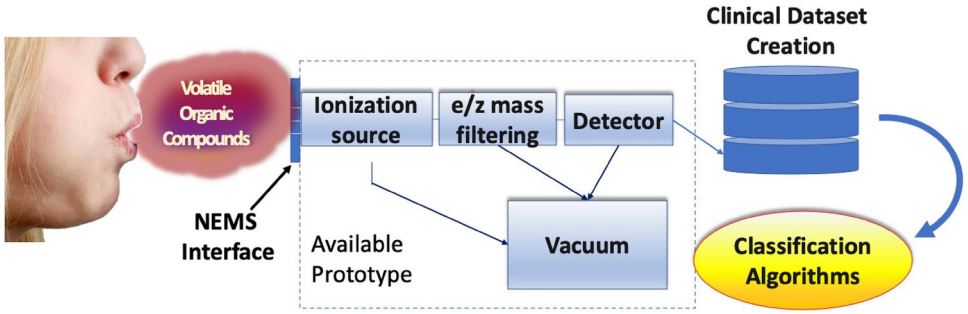

Figure 2. Diagram of the sampling and processing procedure.

by the adopted technology[23,25,26] and also implicitly averages the different phases of the expiratory flow. Then, the filled bag is connected to the inlet valve of the MS apparatus. The inlet valve can have two possible settings: the first setting allows for the sample mixture, at atmospheric pressure, to flow from the bag to the ionization chamber, directly through an original Micro Electro-Mechanical System (MEMS) interface[27]; the second setting connects the MEMS interface to a membrane pump, in order to clean the inlet line, bringing it to vacuum conditions ($\simeq 1e^{-3}\mathrm{mbar}$).A schematic of the sampling system is provided in Fig. 1, and an illustrative diagram of the sample collection process is shown in Fig. 2.

Mass spectra are recorded via a *Varian 1200L* mass analyzer software, which allows the setting of some acquisition parameters like mass range, acquisition time, and electron multiplier (EM) voltage. The latter parameter ultimately sets the detector amplification factor. We recorded mass spectra in the following ranges:

- 10–51 m/z, with an acquisition time of 10 s and EM voltage of 1000 V;
- 49–151 m/z, with an acquisition time of 14 s and EM voltage of 1800 V;
- 149–251 m/z, with an acquisition time of 14 s and EM voltage of 1800 V;
- 249–351 m/z, with an acquisition time of 14 s and EM voltage of 1800 V;To avoid signal saturation, the amplification in the first mass range was reduced due to the presence of the most abundant breath components, namely CO_2 (44 m/z) main peak, N_2 (28 m/z) main peak and O_2 (32 m/z) main peak. For each breath sample, 10–20 acquisitions were taken at fixed time intervals, allowing for the collection of multiple data points from each patient. This approach of acquiring multiple samples per patient enhances the robustness of the mass spectrum analysis by averaging out potential variations and noise in individual measurements. Finally, the raw spectra are filtered and analyzed using our proposed method. The successive analysis of these multiple acquisitions improves the reliability and accuracy of the breath fingerprint profile, leading to more consistent and representative results. This methodology ensures that the breath fingerprint are not anomalies but are reflective of the patient's actual metabolic state, thereby increasing the diagnostic precision of the breath analysis.

The acquisition of each mass range for a subject under test takes less than two minutes (approximately one and a half minutes), requiring about six minutes to acquire all 4 ranges and thus obtain the complete spectrum. Although the approach is not real-time, it is still significantly faster than traditional methods.

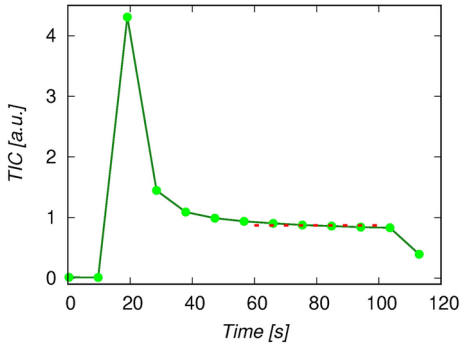

Figure 3. TIC of a recording from one sample. The recording is made up of about 10 acquisitions (green dots), each corresponding to a mass spectrum. The spectra used for the analysis are selected on the plateau of TIC (red dotted region).

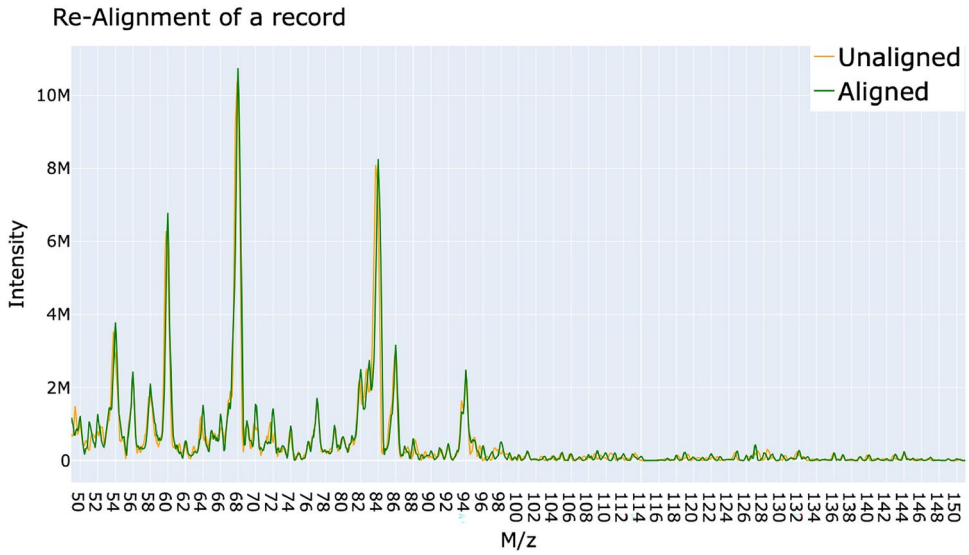

Figure 4. Aligned and non-aligned peaks of the mass spectrum of a single patient.

By summing all the intensities for each m/z in each acquisition, we can obtain the Total Ion Current (TIC) curve plot. Figure 3 shows the TIC behavior when the breath sample flows into the MS system: the initial increase is due to the abrupt pressure change at the valve opening and, after a few tens of seconds, the TIC curve reaches a plateau region[27], when the flux stabilizes. These procedures allowed for storing a dataset composed of the acquisitions of the spectra for each patient.

Pre-processing

Once the raw measures have been obtained, data are cleaned through a pre-processing procedure that reduces noise and machine variation of the acquisitions.

For each acquisition, the recorded m/z positions may be shifted with a specific alignment when the machine records the quantity of the ionized molecules due to measurement noise. A peak-alignment procedure is thus necessary. This procedure enables reducing the noise of the machine and compacting information. The peak alignment procedure is based on moving the peak to the nearest integer position, using them as anchors. The curves between two nearby peaks are stretched or compressed to sustain their original shape, preventing information loss. Stretching and compression between peaks are done by linear interpolation to fit the corresponding segments in the reference. A graphical plot after the peak alignment can be seen in Fig. 4.

As previously mentioned, multiple acquisitions are taken for each patient to ensure the accuracy of the breath analysis. To mitigate potential noise in the measurements and enhance the stability of breath fingerprint recognition, these multiple acquisitions are agglomerated into a single robust mass spectrum. This is achieved by focusing on the plateau zone of the TIC curve, thus the region where the signal stabilizes, indicating that the breath sample flux has reached a steady state.

The plateau zone is composed of the most stable acquisitions, which are indicative of a consistent breath sample. To accurately combine these acquisitions, the first step is to identify the plateau in the TIC curve. A

plateau-searching procedure is implemented, which involves detecting acquisitions that show minimal variation from one another. This is done by computing the gradient of the signal; acquisitions within the plateau zone are those where the gradient is minimal, indicating a stable signal.

Once the plateau zone is identified, the acquisitions within this region are averaged to produce a single, robust mass-spectra measurement. This averaging process reduces the influence of any outlier data points or transient fluctuations, resulting in a more reliable representation of the patient's breath fingerprint. This method enhances the overall robustness and accuracy of the mass spectrometry analysis, ensuring that the VOC profile obtained is both stable and reflective of the true metabolic state of the patient.

The plateau searching procedure was implemented as follows:

- For each acquisition, we computed the gradient of the signals.
- The plateau is defined as a zone that is nearly flat, ideally where the gradient is zero or where the gradient does not vary significantly from zero. To identify this flat zone, we compute a tolerance guard-band, denoted as ϵ, which allows us to classify a region as "flat" if the absolute value of the gradient remains below ϵ. The value of ϵ is determined based on the q-th quantile of the gradient distribution, where q is a parameter within the range $[0, 1]$. This parameter q controls the stringency of the requirement for a constant slope in the plateau region; a lower q value indicates a stricter requirement, leading to a narrower definition of the plateau, while a higher q value allows for more variation in the gradient, resulting in a broader plateau definition.
- The TIC curve may present more than one plateau: the first one is in the region in which the breath sample has not flown yet into the MS machine. This can be composed of 1–3 acquisitions. Thus, to avoid potential errors, we considered the plateau of maximum length, that is, in the region where the ion flow stabilizes.
- Once the plateau of maximum length is found (which varies from 3 to 5 acquisitions for each patient), we computed the standard deviations of acquisitions in this region by deploying a rolling window of size 4. We then chose the 4 acquisitions that minimized the standard deviation, and we computed the mean among these, obtaining a single spectrum for each patient.Computing the average of the 4 acquisitions with minimum standard deviation permits the extraction of a single robust spectrum for each patient.

An alternative approach to artificially increase the dataset is to not average the selected 4 acquisitions but instead insert all of them into the dataset. This approach allows for an increase in both the number of training samples (by a factor of 4) and the variability in the data, which can lead to more accurate models. During the testing phase, instead, we average the acquisitions to obtain a single spectrum for each tested patient.

Some samples may present high noise in the mass spectrum, which can adversely affect the analysis. To address this issue, we identified outlier samples as those with a z-score greater than 8 for at least one feature. Additionally, for some patients, it was not possible to identify a plateau in the TIC curve, leading to their exclusion from the dataset.

To overcome noise in the measurements and possible variations in the machine's settings, a signal filtering and smoothing procedure was applied to the remaining patient samples. The steps involved in this procedure are as follows:

1. **Normalization:** Each spectrum was normalized by dividing by the TIC value to obtain relative information about the breath composition. This step scaled each intensity by the sum of all intensities, bringing the features within the range $(0, 1)$.
2. **Initial High-Pass Filtering:** A high-pass filter was applied, treating as zero any intensity below 0.0001, which was considered noise. This helped in filtering out low-intensity signals that might contribute to noise.
3. **Savitzky–Golay Smoothing and Differentiation:** The Savitzky–Golay Smoothing and Differentiation Filter[28,29] was used to reduce high-frequency noise and align the signals to the baseline. This filter is effective in spectral analysis as it smooths the data while preserving important spectral features.
4. **Secondary High-Pass Filtering:** After smoothing, the high-pass filter was reapplied, treating as zero any intensity below 0.001. This step removed any artifacts that may have been introduced during the smoothing process.The filtering and pre-processing procedures were applied separately to each mass range. Once these steps were completed, the spectra obtained from the 4 mass ranges could be combined to produce a single, comprehensive spectrum spanning the range of 10–351.

If different acquisitions were previously considered for each mass range, merging them involved computing all possible combinations of acquisitions across the mass ranges. This approach effectively augmented the dataset by creating new combinations of the different acquisition spectra for each patient. This procedure can be likened to generating *artificial* patients, where each new patient varies based on one of the 4 segments of the spectrum. An example of the resulting augmented dataset is shown in Table 1.

Finally, the entire spectrum was normalized again by dividing it by the total sum of the intensities, ensuring that only relative information was retained.

Machine-learning models

The mass spectrum analysis was conducted using a comprehensive pipeline comprising several key steps: data normalization, feature selection, dimensionality reduction, and classification. Each stage in this pipeline contributes to the development of a complete ML model. The results of these models are presented in the following sections.

Initially, a Variance Threshold filter was applied to the dataset. This filter removes features with zero variance, thereby eliminating $\mathrm{m/z}$ values for which no intensities were measured post-filtering.

Patient-ID	Acquisitions			
	Mass-Range 1	Mass-Range 2	Mass-Range 3	Mass-Range 4
1-AAAA	1	1	1	1
1-AAAB	1	1	1	2
1-AAAC	1	1	1	3
…				
1-ABCD	1	2	3	4
…				
1-DDDD	4	4	4	4
2-AAAA	1	1	1	1
…				

Table 1. An example of the dataset augmentation procedure. Each row is a pseudo-patient, generated by a particular combination of the different mass-range acquisitions of each actual patient.

Subsequently, each feature was individually normalized using one of two methods: the *Standard Scaler* (SS) or the *Robust Scaler* (RS). The *Standard Scaler* normalizes features by subtracting the mean and scaling according to the variance. In contrast, the *Robust Scaler* subtracts the median and scales based on the interquartile range (the range between the first and third quartiles), which mitigates the impact of outliers.

Further feature reduction was performed to retain only the most informative features. Some of the experiments involved the use of a supervised feature selection method, SURF*, a Relief-based algorithm[30], to select 100 $\mathrm{m/z}$ features. To further reduce dimensionality, Principal Component Analysis (PCA)[31,32] was applied to linearly combine the selected features, utilizing 20 principal components.

Finally, a range of ML classification models was trained to distinguish between COVID-19 positive and negative patients. Various ML techniques have been explored in the context of COVID-19 detection[9,10,15,33,34]. We utilized a combination of state-of-the-art models from[24], including K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), Gradient Boosting (xGB), and Support Vector Machine (SVM) with an *RBF* kernel. Additionally, we implemented an ensemble model that integrates all the aforementioned classifiers using a soft-voting approach (Ens).

In soft voting, predictions from an ensemble of classifiers are amalgamated by considering the probabilities assigned to each class by individual classifiers. The final prediction is determined by selecting the class with the highest cumulative probability across all classifiers involved in the ensemble.

Experimental evaluation

Samples collection

Breath samples were collected from patients and medical personnel at Varese Hospital (Ospedale di Circolo—Fondazione Macchi, ASST Sette Laghi) during an uncontrolled clinical trial conducted as part of a descriptive study. In our sampling setup, humidity and temperature were not directly monitored during sample acquisition. The MS analyzer was stationed in the COVID-19 ward of the hospital, where temperature and humidity were controlled, but specific data on these conditions were not recorded. To minimize variability from uncontrolled factors, all trials and sample collections were conducted in the same room under consistent environmental conditions. No formal sample size evaluation was performed. The volunteers primarily consisted of individuals who suspected they had COVID-19, including those with no symptoms or only mild symptoms.

The acquisition lasted one year, from March 2021 to March 2022, for a total of 302 tested subjects, divided into 91 positive and 211 negative records. The ages of the patients in the study varied from 16 to 88 years. The statistical summary of the ages is as follows:

- Mean age: 55 years
- Standard deviation: 18 years
- Minimum age: 16 years
- 25th percentile: 42 years
- Median age: 57 yearsSome patients have been tested more than once to calibrate the system. The mass spectra have been collected with a *Varian* 1200L mass analyzer, combined with the MEMS interface. As ground truth to confirm SARS-CoV-2 infection, a RT-qPCR nasopharyngeal swab testing has been performed on all subjects. Medical records with detailed information on past health status were available for 176 patients. Among these patients, several had breath-related comorbidities:
-
- 5 Patients had Obstructive Sleep Apnea Syndrome (OSAS).
- 4 Patients had Bronchial Asthma.
- 3 Patients had Sarcoidosis.
- 2 Patients had Pulmonary Fibrosis.
- 13 Patients were diagnosed with Chronic Obstructive Pulmonary Disease (COPD).Additionally, 48 patients had received at least one dose of the COVID-19 vaccine. Respiratory failure was noted in 85 patients, of whom 76 were COVID-19 positive, while the remaining 9 tested negative for the virus. Patients did not undergo physical exercise before the test.[35,36].

The raw dataset presents breath samples for a total of 1,208 acquisitions among the 302 patients. After the outliers' removal procedures and plateau identification, only 287 patients for mass-range 2 and 203 for the whole spectrum 10-351 were retained (Fig. 5). These problems were caused by the highly prototypical nature of the equipment; it is worth noticing that, in a real application, it would have been possible to repeat the measurement.

A graphical view of the spectra resulting from the proposed preprocessing filtering procedure described earlier can be found in Fig. 6 for mass-range 2 and in Fig. 5 for the entire range under analysis.

The data-augmentation procedure, by considering all the different combinations of the acquisitions, leads to the generation of 47,084 samples. A graphical 2D representation of the points obtained with t-SNE dimensionality reduction[37] is presented in Fig. 7. There, a sharp boundary between positive and negative samples can be seen.

Performance evaluation

Patients have been split into training and test data: the training data are used to create the ML models, while the testing data are used only for evaluation purposes. We evaluated the experiments on 10 different training-test splits, averaging the results to obtain an unbiased estimation of the generalization performances of our models. We used a 10-fold stratified cross validation: the acquisitions of each patient are not mixed among training and test sets, to avoid potential information leakages. Thus, each evaluation metric is the mean over 10 run.

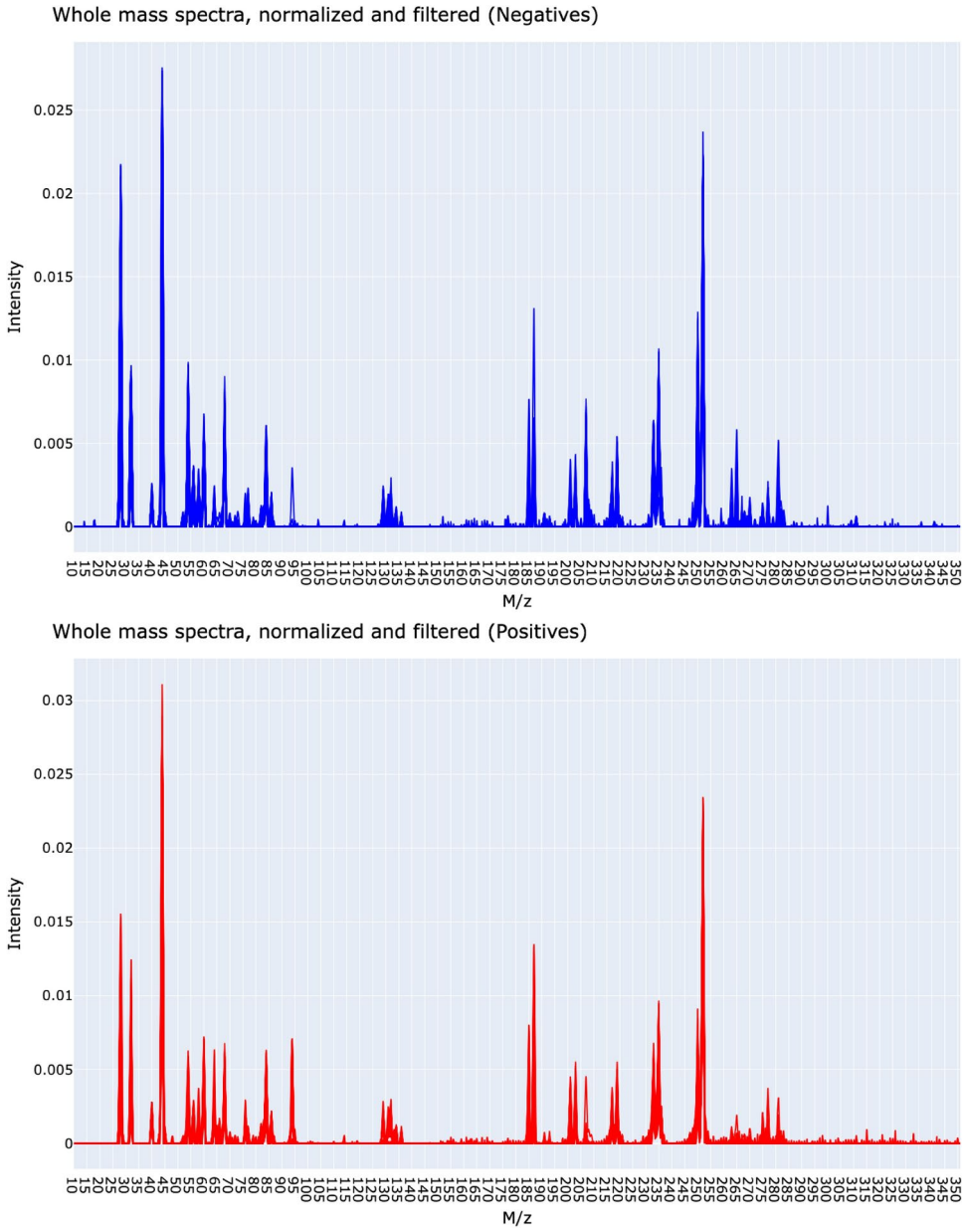

Figure 5. Examples of whole spectra (10–351 m/z) for negative subjects (top, blue) and positive ones (bottom, red).

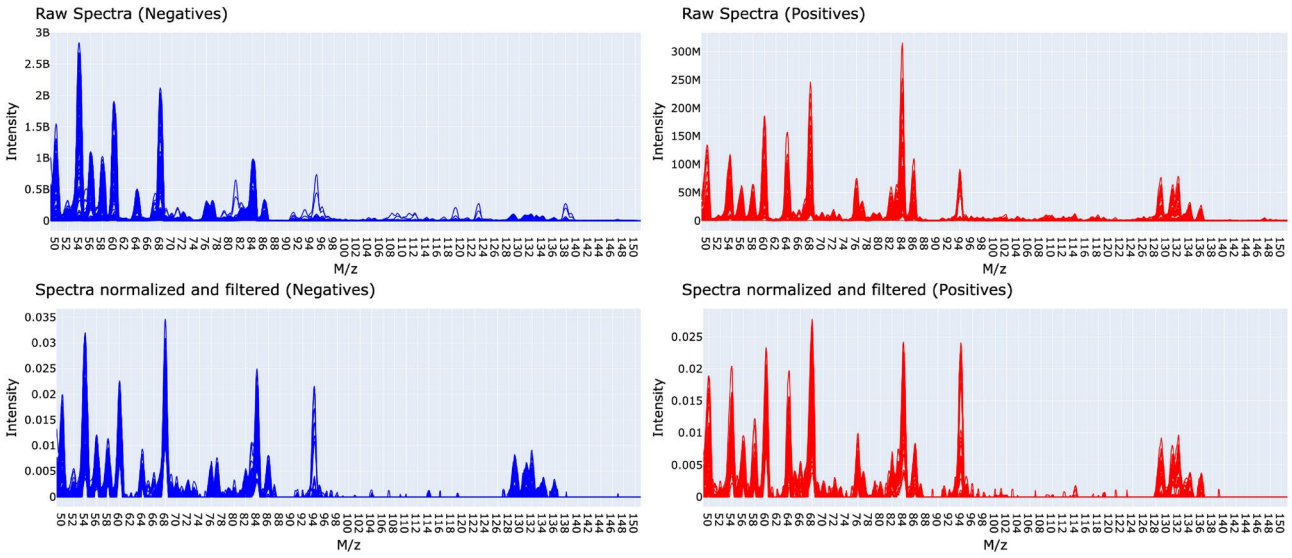

Figure 6. The comparison of spectra before (top) and after the filtering and normalizing procedure (bottom) shows the removal of low-frequency noise. Negative patients are on the left (blue), and positive on the right (red).

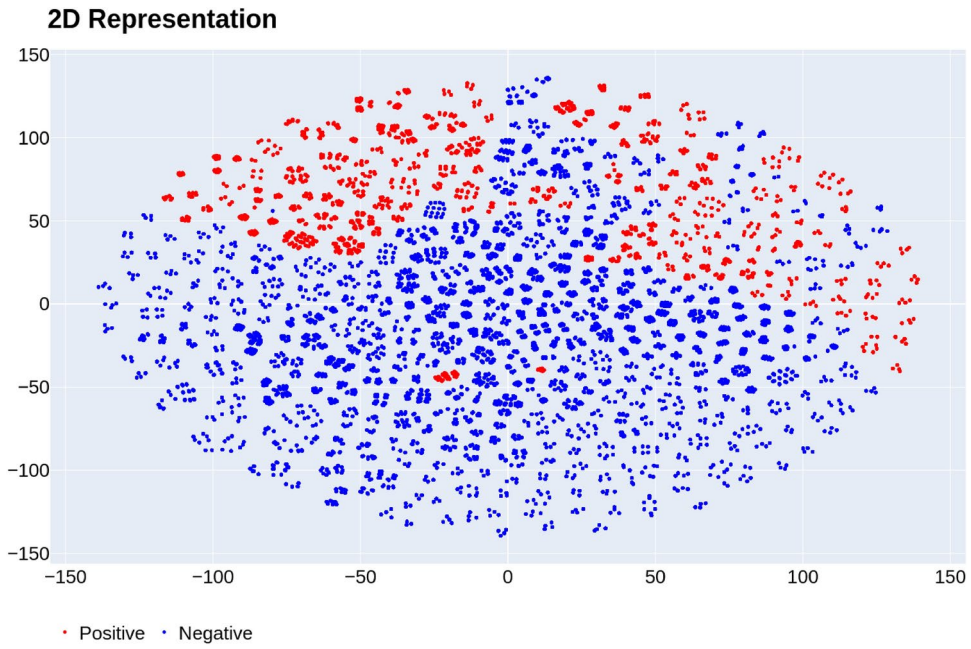

Figure 7. 2D t-SNE representation of the whole spectra for all the 47,084 samples generated.

To address the issue of the high class-imbalance, we utilized a simple oversampling technique of the minority class (i.e., the COVID-19-positive class) in training sets, obtaining the same number of positive and negative samples.

Results are presented in terms of famous classification performances: Balanced Accuracy, Precision, Recall, and F_1-score.

These metrics are computed based on the number of samples correctly and incorrectly predicted by our models.

Precision is the ability of the classifier not to label as positive a sample that is negative, while recall is the ability of the classifier to find all the positive samples.

The balanced accuracy avoids inflated performance estimates on imbalanced datasets. It is the macro-average of recall scores, for each class (the mean of recall for negative and positive classes).

The F_1-score is the harmonic mean of the precision and recall. The relative contribution of precision and recall to the F_1-score are equal. All of these metrics lies in the range [0, 1], where a score equal to 1 (or 100%) means perfect classification performance.

In the experiments on the whole mass spectrum 10–351, with the 4 ranges together, we also computed two additional performance metrics: Specificity and the area under the receiver operating characteristic curve (ROC-AUC). While recall is a measure that evaluates a test's ability to correctly identify unhealthy individuals, specificity carries the same concept but for healthy patients. Specificity is the recall of the negative class. When a test exhibits high specificity, a positive result becomes valuable in confirming the presence of the disease, as the test rarely produces positive outcomes in healthy individuals.

A receiver operating characteristic (ROC) curve is a visual representation that showcases the performance of a binary classifier system as the threshold for classification is adjusted. It plots the true positive rate *(TPR)* against the false positive rate *(FPR)* at different threshold values. *TPR* is also referred to as sensitivity, while *FPR* is the complement of specificity.

The ROC-AUC quantifies the overall performance of the classifier by calculating the area under the ROC curve.

Results

Initially, we trained the models on the different mass ranges separately, without applying any feature pre-processing. Preliminary experiments indicated that mass range 2 (49–151 $\mathrm{m/z}$) was the most effective for classification. The ensemble models achieved an F_1-score of 71% and an accuracy of 84% (Table 2) using 10-fold cross-validation. Subsequently, we applied filtering, normalization, and feature pre-processing and selection specifically to mass range 2. The results from the various trials following these enhancements are summarized in Table 3. The application of the Variance Threshold filter resulted in the retention of 611 $\mathrm{m/z}$ values for Mass Range 2 (Fig. 6) and 1734 features for the entire spectrum covering ranges 1–4 (Fig. 5). This filtering step effectively removes features with zero variance, which do not contribute to the classification task. Selected $\mathrm{m/z}$ in the range 10-351 can be seen in Fig. 8.

It is important to note that we did not perform any compound analysis as the model operates on the raw mass spectrum data. The classification models inherently learns to distinguish between COVID-19 positive and negative classes based on the $\mathrm{m/z}$ values, without needing specific compound identification. This approach ensures that the analysis remains fully transparent to the model, allowing it to directly utilize the spectral data for classification without the need for predefined compound labels.

Filtering, normalizing, and processing the spectra using dimensionality reduction techniques proved beneficial, significantly improving classification performance across all models. For instance, comparing the performance of the Ensemble model before and after pre-processing (Tables 2 and 3), we observed an increase in accuracy from 84 to 87% and an improvement in the F_1-score from 70% to approximately 80% for Mass Range 2. Using the Robust Scaler (RS) instead of the Standard Scaler (SS) further reduced the standard deviation in performance metrics, particularly beneficial for models like SVC, which are sensitive to outliers (Table 3). In contrast, including a Feature Selection step before applying PCA did not yield improvements in performance metrics. The accuracy metrics decreased when using the SURF* algorithm, possibly due to its inability to capture complex feature interactions. Consequently, we found that beyond the simple Variance Threshold filter and PCA, no additional feature manipulation was necessary. On the other hand, PCA consistently enhanced classification performance across all models without the need for prior feature selection. Skipping the feature selection step also helped avoid potential biases in variable selection. Moreover, incorporating multiple acquisitions per patient in the training set and averaging them during testing led to lower prediction errors and higher classification accuracy compared to using a single, robust spectrum (Table 3). This is exemplified by the SVC model: with a single acquisition per patient, the model achieved 76% accuracy and only 57% recall, indicating poor detection of positive patients. However, when using multiple acquisitions, the model's accuracy increased to 89%, and recall improved to 91%. Overall, the SVC and Ensemble models demonstrated the highest recall, effectively identifying true positive patients. In principle, further improvements in prediction performance could be achieved through a fine-tuning of hyperparameters.

The pre-processing steps for the experiments conducted across the entire mass range (10–351, results in Table 4) included a data augmentation procedure, spectra normalization and filtering, the application of a robust

Model	Mass range	Accuracy	Precision	Recall	F_1-Score
RF	1	0.78	0.65	0.56	0.60
RF	2	0.82	0.69	0.69	0.68
RF	3	0.82	0.72	0.57	0.63
RF	4	0.79	0.65	0.53	0.57
Ens	1	0.82	0.70	0.67	0.68
Ens	2	0.84	0.75	0.69	0.71
Ens	3	0.80	0.68	0.57	0.61
Ens	4	0.77	0.62	0.60	0.59

Table 2. Mean Results on test sets (10 splits) with different mass ranges. No features pre-processing.

Alg.	Filtering	Feat. Sel.	Acquisition	Accuracy	Precision	Recall	F_1-Score
xGB	No	PCA	Single	0.83 ± 0.04	0.74 ± 0.10	0.77 ± 0.08	0.75 ± 0.06
xGB	Yes (SS)	PCA	Single	0.88 ± 0.07	0.79 ± 0.08	0.86 ± 0.13	0.82 ± 0.09
xGB	Yes (SS)	PCA	Multiple	0.93 ± 0.05	0.85 ± 0.14	0.94 ± 0.07	0.88 ± 0.09
xGB	Yes (SS)	SURF*, PCA	Multiple	0.86 ± 0.07	0.78 ± 0.14	0.83 ± 0.10	0.80 ± 0.10
xGB	Yes (RS)	PCA	Multiple	0.90 ± 0.04	0.82 ± 0.10	0.89 ± 0.08	0.84 ± 0.05
KNN	No	PCA	Single	0.89 ± 0.05	0.73 ± 0.09	0.92 ± 0.08	0.81 ± 0.07
KNN	Yes (SS)	PCA	Single	0.87 ± 0.05	0.73 ± 0.07	0.89 ± 0.08	0.80 ± 0.07
KNN	Yes (SS)	PCA	Multiple	0.91 ± 0.07	0.80 ± 0.14	0.93 ± 0.10	0.85 ± 0.11
KNN	Yes (SS)	SURF*, PCA	Multiple	0.85 ± 0.07	0.70 ± 0.13	0.87 ± 0.10	0.77 ± 0.11
KNN	Yes (RS)	PCA	Multiple	0.91 ± 0.05	0.78 ± 0.12	0.94 ± 0.08	0.84 ± 0.08
LR	No	PCA	Single	0.86 ± 0.05	0.71 ± 0.12	0.88 ± 0.07	0.78 ± 0.08
LR	Yes (SS)	PCA	Single	0.85 ± 0.06	0.71 ± 0.08	0.85 ± 0.10	0.77 ± 0.07
LR	Yes (SS)	PCA	Multiple	0.88 ± 0.06	0.73 ± 0.14	0.92 ± 0.10	0.80 ± 0.10
LR	Yes (SS)	SURF*, PCA	Multiple	0.88 ± 0.07	0.74 ± 0.13	0.89 ± 0.10	0.80 ± 0.10
LR	Yes (RS)	PCA	Multiple	0.89 ± 0.05	0.76 ± 0.11	0.91 ± 0.08	0.82 ± 0.08
RF	No	PCA	Single	0.87 ± 0.05	0.79 ± 0.10	0.84 ± 0.09	0.81 ± 0.07
RF	Yes (SS)	PCA	Single	0.89 ± 0.04	0.82 ± 0.11	0.87 ± 0.08	0.84 ± 0.05
RF	Yes (SS)	PCA	Multiple	0.91 ± 0.07	0.80 ± 0.14	0.92 ± 0.09	0.85 ± 0.10
RF	Yes (SS)	SURF*, PCA	Multiple	0.85 ± 0.09	0.75 ± 0.16	0.82 ± 0.13	0.78 ± 0.12
RF	Yes (RS)	PCA	Multiple	0.90 ± 0.03	0.81 ± 0.07	0.90 ± 0.07	0.84 ± 0.04
SVC	No	PCA	Single	0.65 ± 0.07	0.82 ± 0.20	0.32 ± 0.13	0.45 ± 0.16
SVC	Yes (SS)	PCA	Single	0.76 ± 0.10	0.82 ± 0.15	0.57 ± 0.17	0.66 ± 0.15
SVC	Yes (SS)	PCA	Multiple	0.89 ± 0.05	0.76 ± 0.13	0.91 ± 0.13	0.81 ± 0.07
SVC	Yes (SS)	SURF*, PCA	Multiple	0.86 ± 0.08	0.77 ± 0.17	0.83 ± 0.09	0.79 ± 0.12
SVC	Yes (RS)	PCA	Multiple	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.85 \pm 0.09}$	$\mathbf{0.94 \pm 0.07}$	$\mathbf{0.89 \pm 0.06}$
Ens.	No	PCA	Single	0.87 ± 0.03	0.77 ± 0.10	0.84 ± 0.06	0.80 ± 0.06
Ens.	Yes (SS)	PCA	Single	0.90 ± 0.07	0.82 ± 0.11	0.89 ± 0.10	0.85 ± 0.09
Ens.	Yes (SS)	PCA	Multiple	0.93 ± 0.07	0.83 ± 0.15	0.94 ± 0.09	0.87 ± 0.11
Ens.	Yes (SS)	SURF*, PCA	Multiple	0.89 ± 0.07	0.79 ± 0.16	0.88 ± 0.08	0.82 ± 0.12
Ens.	Yes (RS)	PCA	Multiple	0.92 ± 0.04	0.81 ± 0.09	0.94 ± 0.07	0.87 ± 0.06

Table 3. Mean Results on tes sets (10 splits) on mass range 2. Models with the highest predictive power are in bold.

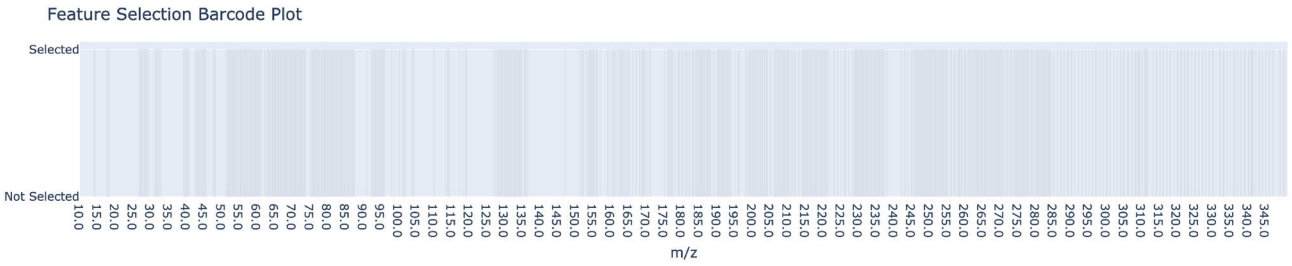

Figure 8. Outcome of feature selection using the Variance Threshold method applied post-filtering, whole mass range 10–351 m/z.

scaler, and PCA for feature extraction. These steps were identified as the most effective in achieving optimal accuracy based on the results of the Mass Range 2 experiments.

This configuration yielded the best performance metrics, as shown in Table 4. Using the ensemble method, we achieved 95% accuracy, 94% recall, 96% specificity, and an F_1-score of 92%.

It is important to note that, although the number of patients under testing was 203, the data augmentation procedure led to a dataset consisting of approximately 47,000 samples.

The code used for spectra analysis, along with the experimental results, has been made available in a public GitHub repository. No data about the patients were released.

Alg.	Accuracy	Precision	Recall	F_1-Score	Specificity	ROC-AUC
KNN	0.93 ± 0.06	0.87 ± 0.09	0.92 ± 0.09	0.89 ± 0.08	0.94 ± 0.04	0.95 ± 0.04
RF	0.91 ± 0.06	0.88 ± 0.10	0.87 ± 0.12	0.87 ± 0.07	0.95 ± 0.04	0.98 ± 0.03
LR	0.94 ± 0.04	0.84 ± 0.12	0.96 ± 0.07	0.89 ± 0.07	0.93 ± 0.05	0.97 ± 0.04
xGB	0.94 ± 0.03	0.88 ± 0.08	0.93 ± 0.07	0.90 ± 0.03	0.95 ± 0.03	0.98 ± 0.03
SVC	0.93 ± 0.06	0.89 ± 0.09	0.90 ± 0.12	0.88 ± 0.06	0.95 ± 0.04	0.98 ± 0.02
Ens.	$\mathbf{0.95 \pm 0.04}$	$\mathbf{0.90 \pm 0.08}$	$\mathbf{0.94 \pm 0.07}$	$\mathbf{0.92 \pm 0.05}$	$\mathbf{0.96 \pm 0.03}$	$\mathbf{0.98 \pm 0.03}$

Table 4. Mean Results on the test sets (10 splits) for the whole mass range 10–351. Filtering the spectra and pre-processing the features (Robust Scaler and PCA) were applied. Models with the highest predictive power are in bold.

Second acquisition campaign

In 2024, we carried out a subsequent round of sample acquisition, acquiring 20 additional negative samples as an independent cohort. Due to challenges in sourcing COVID-19-positive samples, only negative samples were gathered. These new patients were then assessed using the previously developed models. We utilized the ensemble of models trained on the complete mass spectra 10–351, with the data-augmentation procedure. We applied the filtering procedure, resulting in 16 refined spectra. In previous experiments, this approach consistently exhibited better performance. The ensemble model achieved a 100% accuracy rate with the new patients, suggesting promising potential for the proposed methodology despite the presence of time drift. However, additional analyses on a larger patient cohort are necessary to assess the robustness of these findings.

Conclusion

In this study, we presented a comprehensive framework for the detection of COVID-19 using breath samples analyzed by a novel portable MS device based on nanotechnology. This device is capable of analyzing human breath within approximately two minutes, producing a mass spectrum in the range of 10–351 $\mathrm{m/z}$, divided into 4 sub-ranges. Experimental results demonstrated that these mass spectra can be effectively utilized to detect the presence of COVID-19 through ML classification models. A key contribution of our work is the development and implementation of a robust filtering procedure using the Savitzky-Golay filter, which significantly reduces noise in the spectral data. Results indicate that even with relatively simple ML models, we can achieve high classification performance. Specifically, for the mass range of 49–151m/z, which was identified as the most informative for COVID-19 prediction, we achieved an accuracy of approximately 93% and a recall of 94%. The application of robust scaling techniques, which leverage the median and interquartile range for normalization, coupled PCA for feature extraction, further improved our prediction capabilities. By merging all the spectral sub-ranges into a single dataset, we were able to attain even higher classification performance, achieving up to 95% accuracy, 94% recall, and a 98% ROC-AUC. These results underscore the potential of this portable MS machine to be deployed in COVID-19 testing hubs, offering a rapid, non-invasive, and reliable method for detecting the disease. The integration of ML with mass spectrometry opens up promising avenues for the early detection of various diseases, providing rapid test results while minimizing the risk of infection for healthcare providers.

While our study demonstrates the effectiveness of this approach for COVID-19 detection, there are several areas for future research that could further enhance its utility and application. First, expanding the dataset to include a larger and more diverse patient population would help in validating the robustness and generalizability of the proposed model. Second, exploring advanced ML techniques, such as deep learning models, could potentially improve classification performance even further, especially in more complex scenarios involving multiple diseases. Third, we aim to study the reconstruction of VOCs profiles from the mass spectra, to find correlation between specific VOC patterns and the presence of COVID-19. Moreover, investigating the application of this framework to other respiratory conditions or infectious diseases could broaden the scope of its applicability. Future research could also focus on real-time analysis and decision-making capabilities, integrating this technology into telemedicine platforms to facilitate remote diagnostics. Finally, further refinement of the MS device, including its portability and ease of use, could accelerate its adoption in clinical settings, contributing to more widespread and accessible disease detection.

Data availability

The research data from this study has been anonymized to protect participant privacy. This data can be made available to qualified researchers upon request, although informed consent for public release was not obtained. Developed code has been made available in a public GitHub repository.

Received: 15 May 2024; Accepted: 23 September 2024
Published online: 06 October 2024

References

1. Chu, D. K. *et al.* Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: A systematic review and meta-analysis. *Lancet (London, England)* **395**, 1973–1987. https://doi.org/10.1016/S0140-6736(20)31142-9 (2020).

2. Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A. & Kozlakidis, Z. Considerations for diagnostic COVID-19 tests. *Nat. Rev. Microbiol.* **19**, 171–183. https://doi.org/10.1038/s41579-020-00461-z (2021).
3. Feng, W. *et al.* Molecular diagnosis of COVID-19: Challenges and research needs. *Anal. Chem.* **92**, 10196–10209. https://doi.org/10.1021/acs.analchem.0c02060 (2020).
4. Eissa, S. & Zourob, M. Development of a low-cost cotton-tipped electrochemical immunosensor for the detection of SARS-CoV-2. *Anal. Chem.* **93**, 1826–1833. https://doi.org/10.1021/acs.analchem.0c04719 (2021).
5. Bahreini, F., Najafi, R., Amini, R., Khazaei, S. & Bashirian, S. Reducing false negative PCR test for COVID-19. *Int. J. Matern. Child Health AIDS (IJMA)* **9**, 408–410. https://doi.org/10.21106/ijma.421 (2020).
6. Wang, K., Zhu, X. & Xu, J. Laboratory biosafety considerations of SARS-CoV-2 at biosafety level 2. *Health Secur.* **18**, 232–236. https://doi.org/10.1089/hs.2020.0021 (2020).
7. Grassin-Delyle, S. *et al.* Metabolomics of exhaled breath in critically ill covid-19 patients: A pilot study. *EBioMedicine* **63**, 103154. https://doi.org/10.1016/j.ebiom.2020.103154 (2021).
8. Sawano, M., Takeshita, K., Ohno, H. & Oka, H. Rt-pcr diagnosis of covid-19 from exhaled breath condensate: A clinical study. *J. Breath Res.*[SPACE]https://doi.org/10.1088/1752-7163/ac0414 (2021).
9. Binson, V. A., Subramoniam, M. & Mathew, L. Prediction of lung cancer with a sensor array based e-nose system using machine learning methods. *Microsyst. Technol.*[SPACE]https://doi.org/10.1007/s00542-024-05656-5 (2024).
10. Binson, V. A. *et al.* Detection of early lung cancer cases in patients with copd using enose technology: A promising non-invasive approach. In *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 1–4. https://doi.org/10.1109/RASSE60029.2023.10363510 (2023).
11. Binson, V. A., Subramoniam, M. & Mathew, L. Detection of copd and lung cancer with electronic nose using ensemble learning methods. *Clin. Chim. Acta* **523**, 231–238. https://doi.org/10.1016/j.cca.2021.10.005 (2021).
12. Lamote, K. *et al.* The scent of COVID-19: Viral (semi-)volatiles as fast diagnostic biomarkers?. *J. Breath Res.* **14**, 042001. https://doi.org/10.1088/1752-7163/aba105 (2020).
13. Song, J.-W. *et al.* Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. *Cell Metab.* **32**, 188-202.e5. https://doi.org/10.1016/j.cmet.2020.06.016 (2020).
14. Tartaglione, E., Barbano, C. A., Berzovini, C., Calandri, M. & Grangetto, M. Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data. *Int. J. Environ. Res. Public Health*[SPACE]https://doi.org/10.3390/ijerph17186933 (2020).
15. Wang, L. *et al.* Artificial intelligence for covid-19: A systematic review. *Front. Med.*[SPACE]https://doi.org/10.3389/fmed.2021.704256 (2021).
16. Ma, P. *et al.* Non?invasive exhaled breath diagnostic and monitoring technologies. *Microw. Opt. Technol. Lett.* **65**, 1475–1488. https://doi.org/10.1002/mop.33133 (2023).
17. Devillier, P. *et al.* Detection of covid-19 through breath sample analysis with an electronic nose. *Revue des Maladies Respiratoires* **41**, 213–214. https://doi.org/10.1016/j.rmr.2024.01.065 (2024).
18. Giovannini, G., Haick, H. & Garoli, D. Detecting COVID-19 from breath: A game changer for a big challenge. *ACS Sens.* **6**, 1408–1417. https://doi.org/10.1021/acssensors.1c00312 (2021).
19. Zhang, P. *et al.* A feasibility study of COVID-19 detection using breath analysis by high-pressure photon ionization time-of-flight mass spectrometry. *J. Breath Res.* **16**, 046009. https://doi.org/10.1088/1752-7163/ac8ea1 (2022).
20. Liu, X. *et al.* Association of volatile organic compound levels with chronic obstructive pulmonary diseases in nhanes 2013–2016. *Sci. Rep.* **14**, 16085. https://doi.org/10.1038/s41598-024-67210-7 (2024).
21. Khamas, S. S., Bahmani, A. H. A., Vijverberg, S. J. H., Brinkman, P. & Zee, A.H.M.-vd. Exhaled volatile organic compounds associated with risk factors for obstructive pulmonary diseases: A systematic review. *ERJ Open Res.*[SPACE]https://doi.org/10.1183/23120541.00143-2023 (2023).
22. Binson, V., Akbar, R., Thankachan, N. & Thomas, S. Design and construction of a portable e-nose system for human exhaled breath voc analysis. *Mater. Today Proc.* **58**, 422–427, https://doi.org/10.1016/j.matpr.2022.02.388 (2022). International Conference on Artificial Intelligence & Energy Systems.
23. Franceschelli, L. *et al.* Real-time gas mass spectroscopy by multivariate analysis. *Sci. Rep.* **13**, 6059. https://doi.org/10.1038/s41598-023-33188-x (2023).
24. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
25. Mensa, G. & Correale, R. Portable electronic device for the analysis of a gaseous composition (2019).
26. Mensa, G. & Correale, R. Portable electronic system for the analysis of time-variable gaseous flows (2017).
27. Bagolini, A., Correale, R., Picciotto, A., Di Lorenzo, M. & Scapinello, M. Mems membranes with nanoscale holes for analytical applications. *Membranes*[SPACE]https://doi.org/10.3390/membranes11020074 (2021).
28. Savitzky, A. & Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639. https://doi.org/10.1021/ac60214a047 (1964).
29. Gallagher, N. B. Savitzky-golay smoothing and differentiation filter. *Eigenvector Research Incorporated* (2020).
30. Kira, K. & Rendell, L. A. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, 129–134 (AAAI Press, 1992).
31. Gao, Z., Zhang, W., Li, J., Zhang, J. & Huang, G. Principal component analysis-based feature selection for machine learning: A review. *Expert Syst. Appl.* **140**, 112920 (2020).
32. Jolliffe, I. T. *Principal Component Analysis*, vol. 2 (Wiley Online Library, 2002).
33. Costa, Y. M. G. *et al.* COVID-19 detection on chest x-ray and ct scan: A review of the top-100 most cited papers. *Sensors*[SPACE]https://doi.org/10.3390/s22197303 (2022).
34. Smolinska, A. *et al.* Current breathomics-a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J. Breath Res.* **8**, 027105. https://doi.org/10.1088/1752-7155/8/2/027105 (2014).
35. Bikov, A. *et al.* Exercise changes volatiles in exhaled breath assessed by an electronic nose. *Acta Physiol. Hung.* **98**, 321–328 (2011).
36. Bikov, A. *et al.* Exhaled breath condensate ph decreases during exercise-induced bronchoconstriction. *Respirology* **19**, 563–569 (2014).
37. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

Acknowledgements

The authors wish to thank Matteo Serra for his contribution.

Author contributions

N. B., R. Ca., M. C., R. Co., and G. S. defined the research goal, developed the methodology, designed the study, analyzed the data, and drafted the manuscript; M. C. and R. Co. focusing more on the hardware, N. B., R. Ca., and G. S., more one the software. F. C., C. C., D. D. G., F. D., G. P., P. S., and S. T. supervised the data collection and contributed to data analysis and interpretation. All authors reviewed the manuscript, gave their final approval, and agreed to be accountable for all aspects of the work.

Declarations

Competing Interests
The authors declare no competing interests.

Ethics approval
The experiments involving the collection of breath samples from patients and medical personnel at Varese Hospital (Ospedale di Circolo—Fondazione Macchi, ASST Sette Laghi) were conducted in accordance with ethical principles and guidelines. Informed consent was obtained from all participants involved in the study, including patients and medical personnel prior to their inclusion in the research. Participants voluntarily provided their consent to participate.

Additional information
Correspondence and requests for materials should be addressed to N.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.