scientific reports

OPEN

Machine learning allows robust classification of lung neoplasm tissue using an electronic biopsy through minimally-invasive electrical impedance spectroscopy

Georgina Company-Se[1], Virginia Pajares[2], Albert Rafecas-Codern[2], Pere J. Riu[1], Javier Rosell-Ferrer[1], Ramon Bragós[1] & Lexa Nescolarde[1]✉

New bronchoscopy techniques like radial probe endobronchial ultrasound have been developed for real-time sampling characterization, but their use is still limited. This study aims to use classification algorithms with minimally invasive electrical impedance spectroscopy to improve neoplastic lung tissue identification during biopsies. Decision Tree, Support Vector Machines (SVM), Ensemble Method, K-Nearest Neighbors, Naïve Bayes and Discriminant Analysis were applied using mean averaged bioimpedance modulus and phase angle spectra from lung tissue across 15 frequencies (15–307 kHz). Mann-Whitney U test assessed statistical significance between neoplasm and other tissues. Grid search analysis was conducted to determine the optimal hyperparameter configuration for each model, employing a 5-fold cross-validation approach. Model performance was evaluated using Receiver Operating Characteristic curves, with the Area Under Curve (AUC), precision, recall, and F1-score calculated. All the frequencies used to train and test the algorithms obtained high significant differences between neoplasm and the other types of tissues ($P < 0.001$). All the algorithms implemented obtained an accuracy, AUC and F1-score above the 95% except for Naïve Bayes. Decision Tree, Discriminant Analysis and SVM algorithms are suitable for the implementation of a new low-cost guidance method during bronchoscopy.

Keywords Classification, Machine learning, Minimally-invasive bioimpedance, Bronchoscopy, Neoplasm

Diagnosing lung diseases often requires tissue characterization, necessitating lung samples for an accurate final diagnosis. Current imaging methods can guide the diagnosis but are limited as they do not provide real-time guidance for collecting lung tissue samples. The studies found in literature[1,2] lack in consistency regarding the sensitivity of the biopsy in bronchoscopy procedures for lung neoplasm tissue sampling. The studies found state that the sensitivity for this procedure is between 60% and 88%. In addition, another factor for sensitivity variation would be the biopsy method implemented. Recently, new bronchoscopy techniques like radial probe endobronchial ultrasound and electromagnetic navigation bronchoscopy have been developed for real-time sampling characterization. However, their high cost limits availability in Interventional Pulmonology units, and their use remains suboptimal[3,4]. Table 1 summarizes the pros and cons of the existing methods for tissue sampling.

To enhance current diagnostic methods, affordable real-time guidance techniques are needed for tissue sampling. Combining electrical impedance spectroscopy (EIS) with artificial intelligence algorithms could enable electronic biopsies, allowing classification and identification of neoplasm tissue. This would ensure accurate tissue sampling and reduce the occurrence of negative and unnecessary biopsies.

Bioimpedance (Z) is the tissue's opposition to electrical current flow, which varies with frequency when alternating current is applied. This measurement across a wide range of frequencies is known as electrical impedance spectroscopy (EIS). Bioimpedance consists of resistance (R), related to extracellular and intracellular mediums, and reactance (Xc), related to cell membrane capacitance. From these, two parameters are derived:

[1]Department of Electronic Engineering, Universitat Politècnica de Catalunya, Barcelona 08034, Spain. [2]Department of Respiratory Medicine, Hospital de la Santa Creu i Sant Pau, Barcelona 08041, Spain. ✉email: lexa.nescolarde@upc.edu

Method	Pros	Cons
Imaging methods (x-rays, nuclear medicine)	• Characterization, such as its size, shape, density and metabolism of lesions • Location of lesions • Guide biopsy by CT (tomography scan)	• Not real-time guidance • Exposure to radiation or radioactive substances • False positive in benign lesions • False negatives in malignant lesions • High cost of nuclear medicine
Bronchoscopy (conventional diagnostic bronchoscopy)	• Direct visualization allows characterization, location and extent of tumor • Direct tissue sampling by endobronchial biopsy	• Variations in sensitivity depending on biopsy method • Potential complications (bleeding, pneumothorax) • Limitations to visualize and to reach most peripheral lesions • False negative results in malignant lesions
New bronchoscopy techniques (navigation, virtual)	• Real-time sampling characterization • Increase diagnostic yield • Less risks of pneumothorax and bleeding • Increase range of accuracy in most peripheral lesions • Allows mark lesions before performing a surgery	• **Navigation**: High cost, Time may take longer than conventional bronchoscopy and Dependence on operator experience • **Virtual**: not allow real-time sampling, is a planning tool, not a biopsy procedure itself. It requires conventional or post-navigation bronchoscopy for sampling and dependency on the quality of the CT

Table 1. Summary of the pros and cons of the existing methods for tissue sampling.

the impedance modulus (|Z|) defined as $\sqrt{R^2 + Xc^2}$ and the phase angle (PA) defined as $\tan^{-1}(\frac{Xc}{R})$. At low frequencies, current flows through the extracellular medium only, while at high frequencies, it penetrates cell membranes, flowing through both intra- and extracellular mediums[5–7]. Thus, changes in bioimpedance values are expected depending on tissue characteristics.

Previous studies have explored using bioimpedance for lung tissue characterization[8–12]. Meroni et al.[8]. developed an impedance meter for living tissues to test if electrical impedance spectroscopy was helpful in providing information about the structure and the properties of tissues. To validate the instrument, they performed ex-vivo impedance measurements in 3 different rabbits from 6 different tissue types finding statistical significance for the discrimination among the multiple tissues. Toso et al.[9]. evaluated the distribution of the impedance vectors obtained at 50 kHz of frequency from 63 adult male patients with lung cancer and compared the results against 56 healthy subjects obtaining significant differences between cancer patients and control subjects due to a reduced reactance component. Baarends et al.[10]. predicted total body water (TBW) and extracellular water (ECW) in patients with chronic obstructive pulmonary disease (COPD) using bioelectrical impedance spectroscopy (BIS). They concluded that predicted TBW using BIS was comparable to actual TBW, but presented no improvement of the prediction of TBW using bioelectrical impedance analysis (BIA) at 50 kHz. They also found that prediction of ECW had still limitations. These three previous studies focused on the application of impedance measurements for other applications different than our study. Regarding the application of electrical impedance spectroscopy for lung neoplasm differentiation two studies are found. Baghbani et al.[11]. , constructed an electrical bioimpedance sensor with a biopsy forceps shape for measuring electrical conductivity of the tissue inside the body. They obtained and verified the relation between electrical conductivity of the tissue and measured electrical potential with COMSOL software. In addition, they designed and experimentally validated a prototype of the sensor. Furthermore, they measured the impedance of pulmonary tissues in three different samples of tissue founding that the sensor could be potentially beneficial to discriminate tumoral tissues from healthy ones in biopsy process. Baghbani et al.[12]. introduced a method to localize in-depth pulmonary nodules intraoperatively by building a bioimpedance probe with four spherical electrodes. They collected in-vitro bioimpedance data of 286 lung tissue samples and applied principal component analysis (PCA) followed by classification algorithms (support vector machine (SVM), linear discriminant analysis (LDA), and K-nearest neighbors (KNN)) to localize the pulmonary nodules by the bioimpedance spectrum of the lung tissue.

Apart from the last two studies introduced, to the extent of the author's knowledge no studies have applied minimally invasive EIS for lung tissue differentiation, except those by their research group. The first one from Baghbani et al.[11]. , obtained the measures from a biopsy sample while the second applied the classification algorithms from in-vitro samples. Measurements performed by our research group consist on measuring the lung samples directly, during a bronchoscopy process to help in sampling location before performing a biopsy. The first study of our research group was performed by Sanchez et al.[13]. where a bioimpedance device was designed and validated for performing minimally-invasive bioimpedance measurements through bronchoscopy.

Later, research focused on validating the best electrode configuration, implementing a calibration method to reduce data variability, and statistically differentiating between lung tissue types[14–16]. To the extent of the authors knowledge there are not current reports describing the application of Machine Learning classification algorithms to classify neoplasm lung tissue by using electrical impedance spectroscopy measurements for the implementation of an electronic biopsy measurement method to complement the actual guidance systems for a bronchoscopy procedure.

The application of ML algorithms for clinical applications has raise importance in recent years. They present an opportunity to predict outcomes and develop new methods of diagnostic as well as improve prognostics[17]. Current studies regarding the implementation of Artificial Intelligence in pneumology, different studies[18–21] performed in lung cancer applying classification algorithms, focused on the early diagnosis of the disease by using genetic data, computed tomography images and dosimetric features. The majority of the studies are focused on the disease prediction based on medical images.

With the above mentioned, the aim of this study is to compare different Machine Learning algorithms for neoplasm lung tissue classification using minimally-invasive electrical impedance spectroscopy measurements for the implementation of a low-cost guidance method during bronchoscopy aiding in precise biopsy region detection.

Materials and methods

Participants

Minimally invasive EIS measurements were performed between November 2021 and August 2022 in 102 patients (Age: 66 ± 14 year; Weight: 74.5 ± 17.2 kg; BMI: 26.8 ± 4.3 kgm-2) with a bronchoscopy prescribed at the "Hospital de la Santa Creu i Sant Pau" of Barcelona. A total number of 116 samples were obtained divided in 29 samples of lung tissue neoplasm and 87 samples of other lung tissue types (emphysema ($N=23$), healthy lung tissue ($N=30$), pneumonia ($N=22$) and fibrosis ($N=12$)).

EIS measurements

Minimally-invasive EIS using the 3-electrode method bioimpedance measurements are obtained through the injection of a multisine current signal (from 1 kHz to 1000 kHz) between a distal tetrapolar catheter electrode and a skin electrode during a bronchoscopy procedure. The voltage induced by the injected current is measured between the distal electrode and a second skin electrode. Impedance signal acquisition time was 12 s using a sample frequency of 60 spectra per second. The complete description of the impedance device as well as the calibration of the measurements is at Company-Se et al.[15]. Measurements with abnormally large impedance values due to lose of contact between the catheter electrodes and the tissue samples were discarded for the analysis. Radiological images (CT or PET/CT) are taken in each patient before bronchoscopy following the diagnostic process. The catheter used to obtain the bioimpedance data is inserted through the working channel of the bronchoscope. Patients are placed in a supine position during the bioimpedance acquisition with the upper airways anaesthetized. Moreover, intravenous sedation is also provided. Biopsy was obtained to confirm the neoplasm diagnosis. Prior to the bioimpedance and biopsy acquisition saline solution is injected to clean and homogenize the tissue conditions among the different patients.

Ethical statement

Ethics approval was obtained from the Hospital de la Santa Creu i Sant Pau (CEIC-73/2020) according to principles of the Declaration of Helsinki for experiments with human beings. The patients/participants provided their written informed consent to participate in this study.

Variables included in the study

The data features used to apply the classification algorithms is constituted by the 12 s mean averaged spectra of the bioimpedance |Z| and PA obtained from 15 frequencies ranged from 15 kHz to 307 kHz. Given the 29 samples of neoplasm and the 87 samples of other lung tissue types, together with the 15 measures of |Z| and the 15 measures of PA, the total dimensionality of the original dataset is 116 samples * 30 features.

Data preprocessing

Synthetic Minority Oversampling Technique (SMOTE) is applied twice[22]. First, SMOTE is applied to balance the dataset. From the 29 original neoplasm samples we created 58 synthetic neoplasm cases with a final first step sample size of 87 samples from lung neoplasm and 87 samples from other types of tissue. SMOTE is again applied to increase the sample size in a 50% more, thus obtaining from the 174 samples a final size of 262 samples (adding 44 synthetic samples per each class). Synthetic data creation has been performed using 5 Nearest Neighbors. After data augmentation, the total dimension of the dataset is 262 samples * 30 features.

Data analysis

The normality of all the features was assessed using the Kolmogorov-Smirnov test. Variables, non-parametric distributed, are described as median (interquartile range, IQR) and (minimum – maximum). Mann-Whitney U test, was used to assess non-normally distributed statistical significance between neoplasm tissue and the other group. The statistical significance was set as $P<0.05$.

Classification models

Decision Tree, Support Vector Machines (SVM), Ensemble Method, K-Nearest Neighbors (KNN), Naïve Bayes and Discriminant Analysis classifiers were evaluated over the impedance modulus and phase angle of 15 frequencies distributed between 15 kHz and 307 kHz. Each dataset was normalized, as the range of values of the impedance modulus is different than the range of values of the impedance phase angle, in order to improve model performance and training efficiency. A grid search analysis was conducted to determine the optimal hyperparameter (HP) configuration for each model, employing a 5-fold cross-validation approach, which has been commonly utilized in prior medical research[23,24]. During every cross-validation, each dataset was partitioned into training (≈ 80% of the data) and test (≈ 20% of the data) and the model was trained and evaluated with each set of partitions. The HP optimized and the range of optimization are described in Table 2.

The software MATLAB version: 23.2.0.2485118 (R2023b), Natick, Massachusetts: The MathWorks Inc.; 2023 has been used to implement the algorithms using the Classification Learner App.

Classification models interpretability and performance assessment

To evaluate the performance of each implemented model, a receiver operating characteristic (ROC) graph was generated. ROC graphs serve as valuable tools for analyzing the effectiveness of classification models by examining their true positive rate in relation to their false positive rate[25]. The diagonal line in a ROC graph represents random guessing, and models positioned below this diagonal are considered less effective than

Model	Hyperparameters
Decision Tree	Maximum Number of Splits = 1 to 261 Split Criterion = Gini's diversity index, Maximum deviance reduction
Support Vector Machines	Box constraint level = 0.001 to 1000 Kernel scale = 0.001 to 1000 Kernel function = Gaussian, Linear, Quadratic, Cubic
Ensemble Method	Ensemble method = Bag, GentleBoost, LogitBoost, AdaBoost, RUSBoost Number of learners = 10 to 500 Learning Rate = 0.001 to 1 Maximum Number of Splits = 1 to 261
K-Nearest Neighbors	Number of Neighbors = 1 to 131 Distance Metric = City block, Chebyshev, Correlation, Cosine, Euclidean, Hamming, Jaccard, Mahalanobis, Minkowski, Spearman Distance weight = Equal, Inverse, Squared Inverse
Naïve Bayes	Distribution names = Gaussian, Kernel Kernel type = Gaussian, Box, Epanechnikov, Triangle
Discriminant Analysis	Discriminant type = Linear, Quadratic, Diagonal Linear, Diagonal Quadratic

Table 2. HP evaluated during the 5-fold cross-validation grid search analysis.

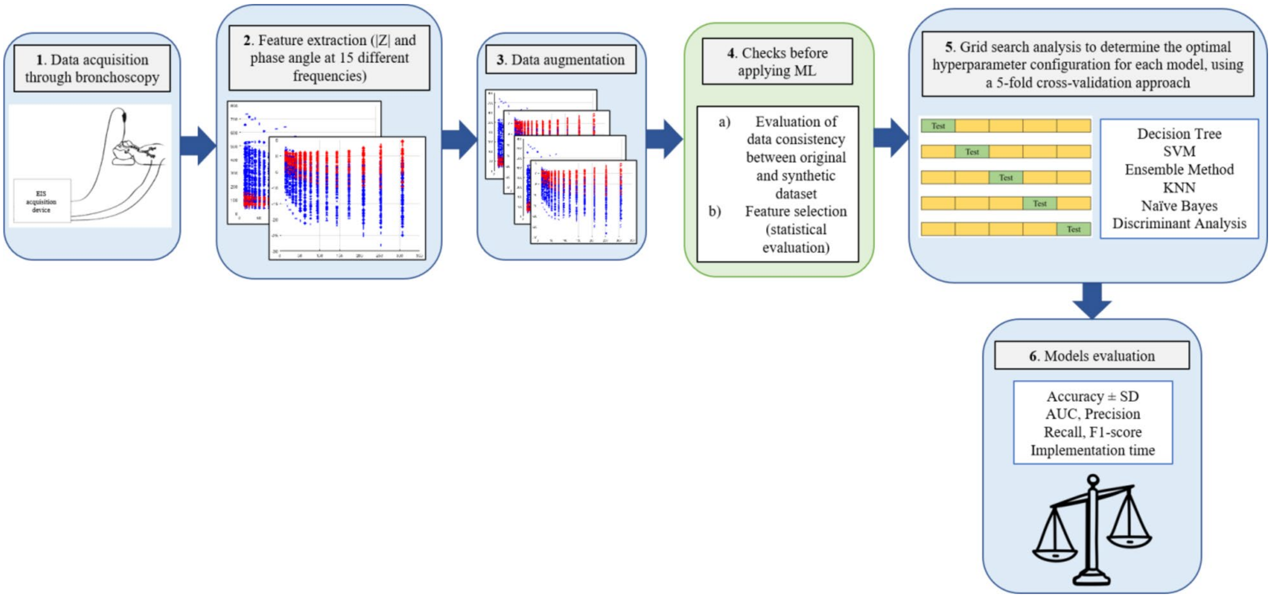

1. Data acquisition through bronchoscopy
EIS acquisition device
2. Feature extraction (|Z| and phase angle at 15 different frequencies)
3. Data augmentation
4. Checks before applying ML
a) Evaluation of data consistency between original and synthetic dataset
b) Feature selection (statistical evaluation)
5. Grid search analysis to determine the optimal hyperparameter configuration for each model, using a 5-fold cross-validation approach
Test
Test
Test
Test
Test
Decision Tree
SVM
Ensemble Method
KNN
Naïve Bayes
Discriminant Analysis
6. Models evaluation
Accuracy ± SD
AUC, Precision
Recall, F1-score
Implementation time

Fig. 1. Schematic diagram showing the process followed from data acquisition to model evaluation for classification to aid in a precise biopsy region detection.

random chance. An ideal classifier is depicted in the top-left corner of the graph, achieving a true positive rate of 1 and a false positive rate of 0. The area under the ROC curve (AUC) is calculated to quantitatively summarize the model's classification performance. In addition to the ROC graph and AUC, the evaluation also includes precision, which is defined as the ratio of correctly identified positive instances (true positives) to the total predicted positive instances (true positives + false positives). Recall, representing the proportion of correctly identified positive instances (true positives) relative to the total actual positive instances (true positives + false negatives), is also considered. Furthermore, the F1-score, a metric that balances precision and recall to provide a comprehensive measure of classification performance, is computed as part of the assessment[26]. Finally, the implementation time for each of the algorithms is also obtained.

The overall process from data acquisition to model evaluation for classification to aid in a precise biopsy region detection is shown in the schematic diagram represented in Fig. 1.

Results

Impedance |Z| and PA mean impedance spectrum

Figures 2 and 3 show the mean impedance spectra of the modulus and phase angle respectively for neoplasm (black), and other tissue (healthy and pathologic, blue) along all the frequencies (15 kHz to 307 kHz) for the original data (left) and for the dataset augmented after SMOTE application (right). The continuous line represents de mean while the dashed and pointed lines represent the ± SD. It represents the evolution of the impedance values (|Z| and PA) with respect to the frequency for both tissue groups.

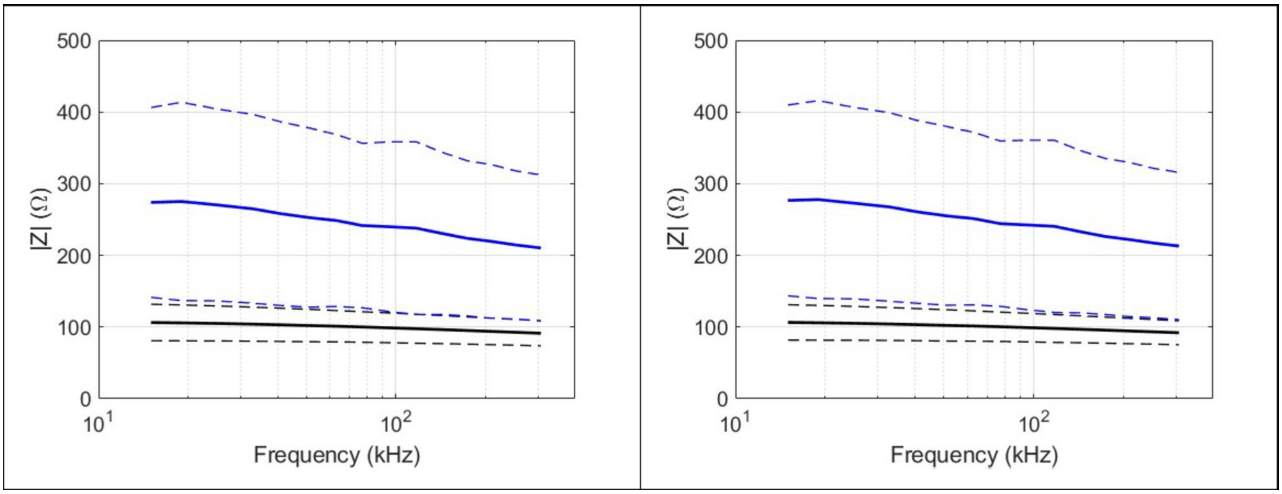


Fig. 2. Mean impedance spectrum of |Z| for original data (left) and for the dataset augmented (right) for all the frequency range. Black: neoplasm; Blue: other types of lung tissue.

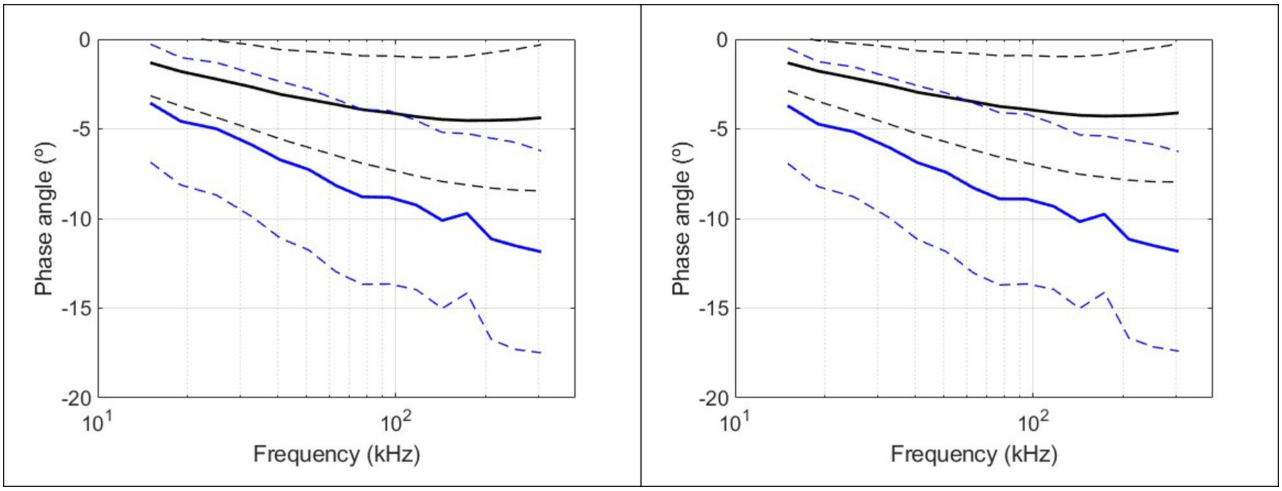


Fig. 3. Mean impedance spectrum of phase angle for original data (left) and for the dataset augmented (right) for all the frequency range. Black: neoplasm; Blue: other types of lung tissue.

Analyzed variables

Table 3 show the descriptive information expressed as median (IQR) (minimum – maximum) of the variables included in the classification models for neoplasm group and for the group that included different types of lung tissue (Other tissue group). In addition, Table 3 shows the statistic U of Mann Whitney and the statistical significance P.

Classification models

Table 4 shows the optimal hyperparameter configuration for each of the classification models obtained from the 5-cross-validation grid search analysis, along with the corresponding test accuracy values and the standard deviation of the accuracy obtained from the cross-validation.

Figure 4 shows the confusion matrices, representing the relationship between actual and predicted classes, for each of the classification models implemented.

Classification models performance assessment

Figure 5 show the ROC curves, that represent a visual representation of the models' performance, obtained for each of the classification models implemented. In addition, the precision, recall and F1-score metrics are also specified.

Table 5 shows the implementation time for each of the algorithms implemented.

F (kHz)	Other tissue	Neoplasm	U	*P*	Other tissue	Neoplasm	U	*P*
	Module (Ω)				Phase angle (°)			
15	278.03 (224.12) (41.53–655.43)	110.06 (36.46) 50.03–156.92)	11.043	< 0.001	-2.97 (3.76) (-13.96–2.65)	-1.39 (1.83) (-5.86–1.54)	-6.851	< 0.001
19	271.82 (222.95) (40.72–736.23)	109.90 (36.30) (50.09–153.96)	10.888	< 0.001	-4.16 (5.17) (-15.90–1.94)	-1.52 (1.94) (-6.12–1.25)	-7.428	< 0.001
25	269.37 (217.50) (40.63–714.23)	109.63 (34.61) (50.12–150.17)	10.895	< 0.001	-4.37 (4.80) (-17.00–1.83)	-1.87 (2.53) (-6.29–0.92)	-7.245	< 0.001
33	260.04 (215.21) (39.84–696.44)	108.94 (33.46) (49.99–146.32)	10.763	< 0.001	-5.39 (5.41) (-18.26–1.45)	-2.32 (3.30) (-7.03–0.98)	-7.643	< 0.001
41	256.99 (207.78) (39.37–703.81)	108.22 (31.88) (49.89–143.08)	10.689	< 0.001	-6.14 (6.03) (-19.72–0.71)	-2.97 (3.77) (-7.48–0.99)	-7.813	< 0.001
51	253.17 (201.67) (38.75–664.60)	107.37 (30.14) (49.72–139.92)	10.604	< 0.001	-6.76 (5.55) (-20.46–0.60)	-3.02 (4.49) (-7.89–1.05)	-8.080	< 0.001
63	251.52 (189.95) (39.03–610.22)	106.56 (29.03) (49.60–136.81)	10.640	< 0.001	-7.26 (6.05) (-21.31–0.18)	-3.34 (5.05) (-8.29–1.14)	-8.599	< 0.001
77	246.26 (178.06) (38.62–549.68)	105.98 (28.39) (49.43–133.80)	10.554	< 0.001	-7.96 (6.29) (-21.61 – (-0.46))	-3.80 (5.13) (-8.71–1.18)	-8.865	< 0.001
95	240.85 (175.05) (37.90–625.97)	105.25 (27.21) (49.18–130.93)	10.437	< 0.001	-7.91 (5.97) (-20.64 – (-0.58))	-3.72 (5.43) (-9.07–1.28)	-8.609	< 0.001
117	235,58 (176.65) (37.18–639.29)	104.38 (25.81) (48.94–128.06)	10.345	< 0.001	-8.37 (5.81) (-20.33 – (-1.33))	-4.15 (5.68) (-9.41–1.34)	-8.853	< 0.001
143	230.76 (160.86) (37.19–569.36)	103.30 (24.43) (48.77–125.33)	10.360	< 0.001	-9.16 (6.25) (-21.96 – (-1.84))	-4.32 (6.01) (-9.71–1.40)	-9.579	< 0.001
173	225.39 (150.72) (36.81–526.93)	102.14 (23.53) (48.60–123.33)	10.327	< 0.001	-9.12 (5.94) (-19.72 – (-1.84))	-4.28 (6.14) (-9.87–1.69)	-9.170	< 0.001
209	216.35 (145.07) (36.34–525.89)	100.35 (22.15) (48.60–123.33)	10.269	< 0.001	-11 (7.64) (-25.14 – (-1.35))	-4.21 (6.25) (-10.10–2.57)	-9.732	< 0.001
253	209.13 (143.51) (36.15–521.80)	97.80 (22.29) (48.34–119.15)	10.300	< 0.001	-11.63 (8.09) (-28.10 – (-1.01))	-4.01 (6.31) (-10.17–3.48)	-9.753	< 0.001
307	202.63 (140.73) (35.63–520.61)	95.37 (20.42) (48.25–116.93)	10.262	< 0.001	-12.50 (7.97) (-25.10 – (-0.56))	-3.93 (6.20) (-10.13–4.61)	-10.034	< 0.001

Table 3. Statistical significance for lung neoplasm differentiation from the other types of tissue of parameters used to train and evaluate the classification models.

Model	Optimal Hyperparameter	Accuracy	SD of Accuracy
Decision Tree	Maxim Number of Splits = 12 Split Criterion = Gini's diversity index	0.966	0.0236
Support Vector Machines	Box constraint level = 999 Kernel scale = 1 Kernel function = Linear	0.981	0.0161
Ensemble Method	Ensemble method = GentleBoost Number of learners = 497 Learning Rate = 0.0014 Maximum Number of Splits = 2	0.969	0.0251
K-Nearest Neighbors	Number of Neighbors = 1 Distance Metric = Chebyshev Distance weight = Inverse	0.958	0.0399
Naïve Bayes	Distribution names = Kernel Kernel type = Box	0.92	0.0504
Discriminant Analysis	Discriminant type = Linear	0.973	0.0157

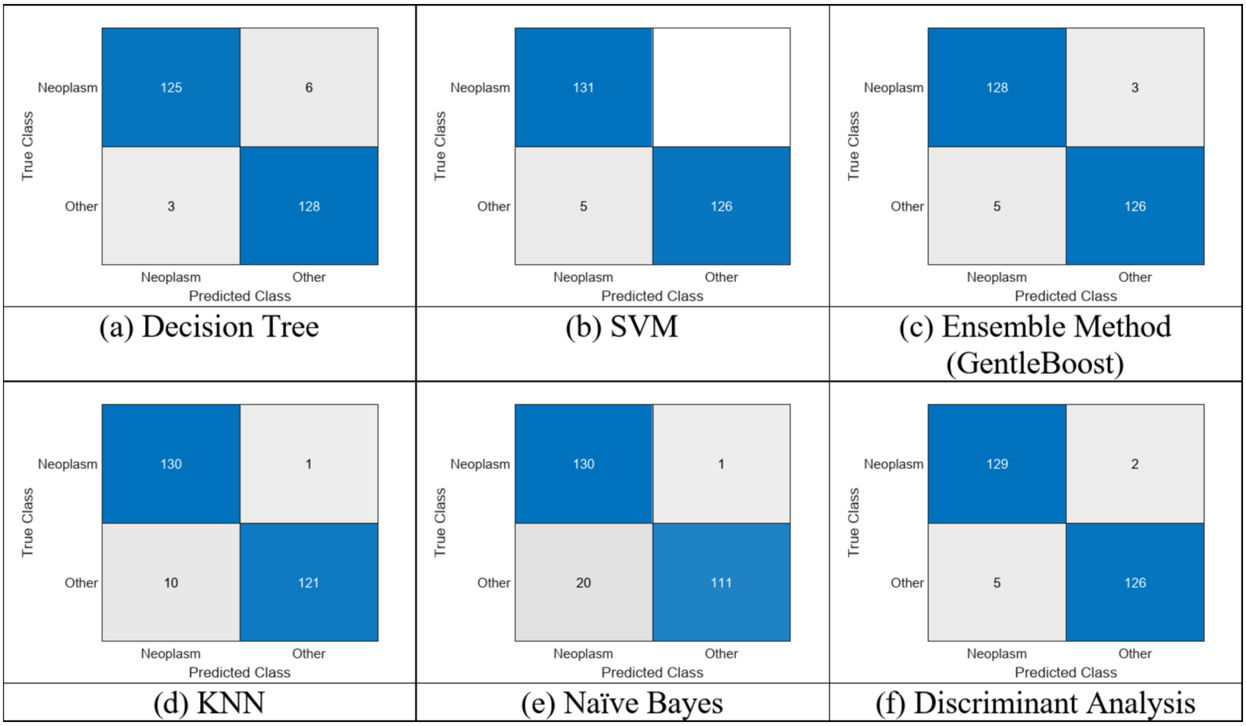
Table 4. Best HP configuration for each model with the accuracy and standard deviation of the accuracy obtained in the 5-fold-cross-validation grid search analysis.

Discussion

The application of Machine Learning classification algorithms are promising tools for the future of the medicine. They present an opportunity to help in the diagnosis of diseases for the tissue characterization.

Different pathologies can affect to the respiratory system such as emphysema, neoplasm, fibrosis or pneumonia. Each of these disorders have their own anatomical and histological changes, thus differences in bioimpedance values are expected[27]. Neoplasm is characterized by an increase in cell concentration as well as tissue vascularization[28] which lowers the module impedance and increases the phase angle with respect to other types of tissue (Figs. 2 and 3).

Regarding data augmentation, large datasets generally enhance classification accuracy, whereas small datasets are prone to overfitting. Data augmentation techniques can mitigate these challenges by generating extra samples



(a) Decision Tree

(b) SVM

(c) Ensemble Method (GentleBoost)

(d) KNN

(e) Naïve Bayes

(f) Discriminant Analysis

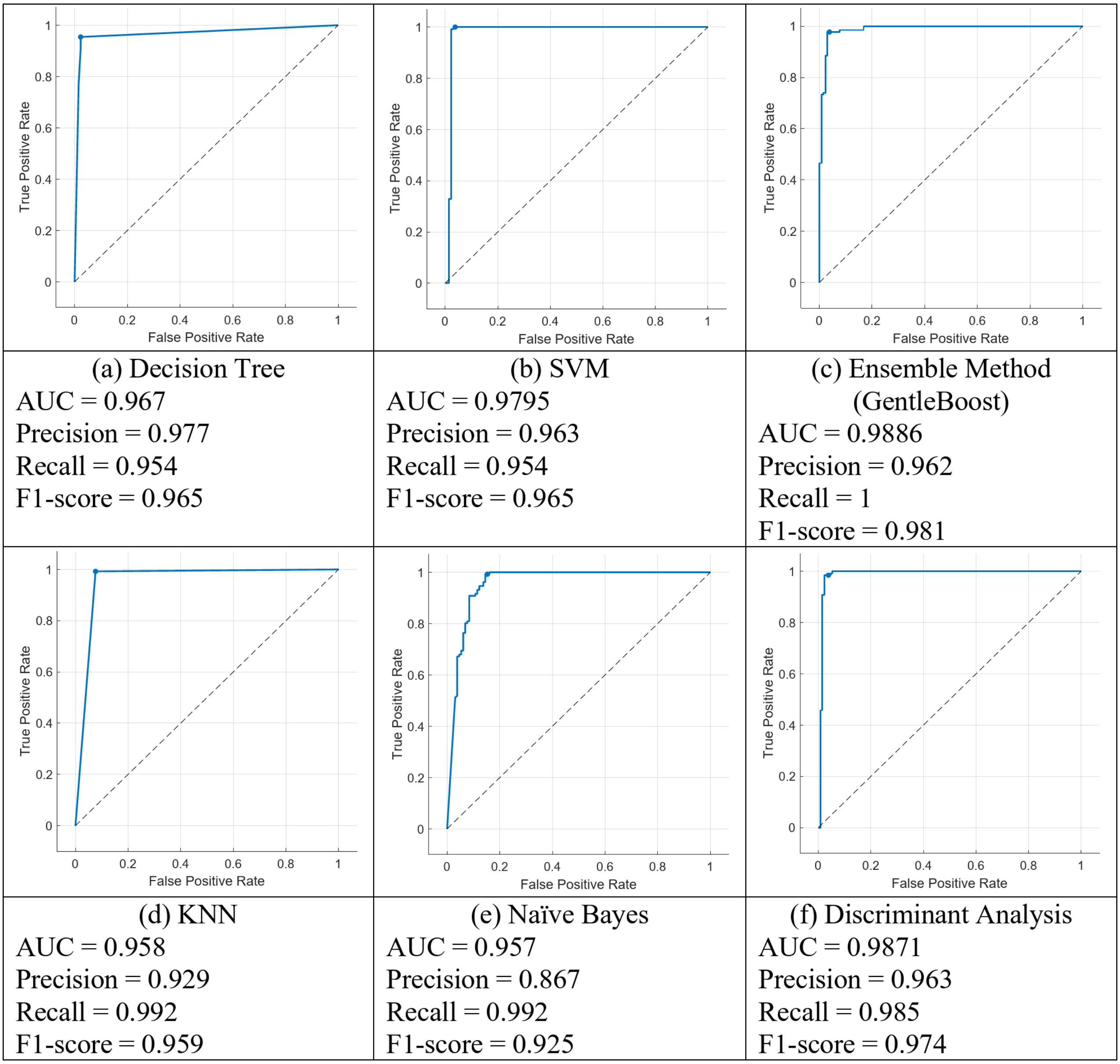
Fig. 4. Confusion matrices for the different algorithms implemented obtained in the 5-fold-cross-validation grid search analysis.

for training models, thereby extracting more meaningful insights from limited data. Researchers leverage these methods to increase dataset diversity[29]. Regarding the application of SMOTE, the new data is created from the average of 5 nearest original samples. According to the results of Figs. 2 and 3, it can be concluded that the algorithm created synthetic data accordingly as the mean impedance spectrum both, for impedance modulus and phase angle does not change from the original data to the dataset with synthetic data.

For the implementation of the classification algorithms, the impedance modulus and phase angle from 15 different frequencies from 15 kHz to 307 kHz have been used. Mann-Whitney U test can only test significance for each of the parameters alone. However, classification algorithms can use all the variable to find complex interconnections among variables to find a robust classification between neoplasm lung tissue or other type of lung tissue. As both, modulus and phase angle have obtained high statistical significance ($P<0.001$) for all the frequencies, and the U statistic does not differ much among variables (Table 3), all the variables have been used for the classification models implementation.

With respect to the classification models, the learning problem consist on finding a complex equation able to classify the samples as better as possible using training data by minimizing the classification error and increasing the accuracy for classification in never seen data. To do that, the different characteristic parameters of each of the algorithms have been optimized with the grid search method. It remains highly effective and ensures the identification of the most optimal combination of model parameters to obtain the higher metrics as possible for the classification task. Thus, the different models implemented are adequately optimized (Table 4). In order to minimize as much as possible the overfitting of the algorithm, validation have been performed by using cross-validation with 5 folds[30]. Regarding the accuracy obtained for each of the classification models implemented, all of them obtained an accuracy higher than 0.95 except for the Naïve Bayes, which has obtained an accuracy of 0.92 in addition of the highest SD of the accuracy (0.05). Furthermore, the HP of Table 4 show that K-Nearest Neighbors used 1 neighbor to classify. This leads to the fact that although the bias of the 1-nearest neighbor estimate is often low, the variance is high[31].

The different studies published regarding the use of classification algorithms for medical applications highlight the high performance of Support Vector Machines algorithm, although there has not been found consistence regarding which algorithm could fit best for medical applications[32–34]. Therefore, it is important to compare the performance of different algorithms to decide which suits best for the application under study. In our case, all of them, except Naïve Bayes and KNN algorithms could serve to detect neoplasm lung tissue through an electronic biopsy, with special focus on Support Vector Machines. Comparing the performance of the implemented algorithms with the study in the field that also applied classification algorithms[12], they found that SVM performed slightly better than LDA and KNN. However, the differences between the performance of the three algorithms are minor. In our case, the differences in performance between SVM and LDA are also minor. However, as the literature suggests, we focused especially on SVM since it seems to be the algorithm that performs the best on medical data. The findings in the literature lead to the conclusion that the application



(a) Decision Tree
AUC = 0.967
Precision = 0.977
Recall = 0.954
F1-score = 0.965

(b) SVM
AUC = 0.9795
Precision = 0.963
Recall = 0.954
F1-score = 0.965

(c) Ensemble Method (GentleBoost)
AUC = 0.9886
Precision = 0.962
Recall = 1
F1-score = 0.981

(d) KNN
AUC = 0.958
Precision = 0.929
Recall = 0.992
F1-score = 0.959

(e) Naïve Bayes
AUC = 0.957
Precision = 0.867
Recall = 0.992
F1-score = 0.925

(f) Discriminant Analysis
AUC = 0.9871
Precision = 0.963
Recall = 0.985
F1-score = 0.974

Fig. 5. ROC curves along with their corresponding AUC and the precision, recall and F1-score obtained in the 5-fold-cross-validation grid search analysis for each model implemented.

Algorithm	Implementation time (seconds)
Decision Tree	38.51
SVM	31.84
GentleBoost	168.88
KNN	36.13
Naïve Bayes	62.26
Discriminant Analysis	39.49

Table 5. Implementation time of the algorithms implemented to classify between neoplasm and health lung tissue.

of already created classification algorithms are useful and safe for new medical applications or to complement already existing medical approaches.

To fully evaluate the performance of the algorithms implemented, the classification report for all the algorithms have been obtained (Fig. 5). We have plotted the ROC curve, used to evaluate the performance of a binary classification method for diagnostic, together with the AUC (Area Under Curve), which is a measure used to evaluate the accuracy of the test[25]. All the classification methods reached an AUC higher than 95% which means the test accuracy is excellent. In order to fully assess the performance of the classification tests, the precision, recall and F1-score have also been calculated. While precision takes more importance when false positives are wanted to be avoided, the recall takes importance when false negative are not desired[35]. F1-score offers a trade-off between precision and recall. For our clinical problem, false positives are not desired, as the biopsies of other tissue rather than neoplasm are not of clinical interest. On the other hand, false negative is also not desired, as we want to accurately localize neoplasm tissue to help in sampling location during bronchoscopy. Therefore, we focus on the F1-score to evaluate the performance of the classification methods implemented.

Given that the classification method is intended for detection of tissue types during bronchoscopy, the duration of bioimpedance data acquisition and processing is critical, with shorter execution times being a significant advantage. The implementation time of the various classification methods employed are described in Table 4 with shorter times in Decision Tree, SVM and Discriminant Analysis algorithms.

According to Fig. 5, except for Naïve Bayes algorithm, all the algorithms obtained an F1-score higher than 0.95, which, together with the accuracy displayed in Table 4 and the execution times shown in Table 5, leads to the conclusion that Decision Tree, SVM and Discriminant Analysis are suitable for our clinical application. In future studies the performance of Decision Tree, Discriminant Analysis and, specially, SVM, will be tested in real-time. In addition, by using an electromagnetic navigation bronchoscopy the usefulness of the classification algorithms to detect peripheral nodules will be evaluated.

Clinical significance

This study demonstrated the usefulness of machine learning algorithms for detecting neoplasm lung tissue during a bronchoscopy by performing an electronic biopsy and measuring the bioimpedance of the tissue. Indeed, thanks to machine learning, healthcare professionals can take advantage of tissue electrical properties variations which may remain unmeasured by actual navigation systems.

In clinical practice, this minimally-invasive sampling localizer could be employed in the interventional pulmonology unit of hospitals for accurate biopsy sampling location during bronchoscopy enabling the decrease the negative biopsies due to sampling errors.

Limitations

Lung nodules, according to their location, can be classified as central nodules and peripheric nodules. During bronchoscopy, patients receive sedation, not anesthesia which implies that movement of the patients are frequent. In order to ensure and minimize the risk of not piercing the pleura, this study only included central nodules, which are analyzed with a conventional bronchoscopy.

Conclusions

The results obtained after the comparison among all the algorithms implemented for neoplasm lung tissue classification using minimally-invasive electrical impedance spectroscopy measurements show that Decision Tree, Discriminant Analysis and Support Vector Machines algorithms, with special emphasis in the last one, are suitable for the implementation of a low-cost guidance method during bronchoscopy and that a new tool could be designed as new guidance tool.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 26 November 2024; Accepted: 17 March 2025
Published online: 21 March 2025

References

1. Andolfi, M. et al. The role of bronchoscopy in the diagnosis of early lung cancer: a review. *J. Thorac. Dis.* **8**, 3329–3337 (2016).
2. Rivera, M. P., Mehta, A. C. & Wahidi, M. M. Establishing the diagnosis of lung cancer: diagnosis and management of lung cancer, 3rd ed: American college of chest physicians Evidence-Based clinical practice guidelines. *Chest* **143**, e142S–e165S (2013).
3. Herth, F. J. F., Eberhardt, R., Becker, H. D. & Ernst, A. Endobronchial ultrasound-guided transbronchial lung biopsy in fluoroscopically invisible solitary pulmonary nodules: a prospective trial. *Chest* **129**, 147–150 (2006).
4. Folch, E. E. et al. Electromagnetic navigation bronchoscopy for peripheral pulmonary lesions: One-Year results of the prospective, multicenter NAVIGATE study. *J. Thorac. Oncol.* **14**, 445–458 (2019).
5. Lukaski, H. C. Biological indexes considered in the derivation of the bioelectrical impedance analysis. *Am. J. Clin. Nutr.* **64**, 397S–404S (1996).
6. Lukaski, H. C., Diaz, V., Talluri, N., Nescolarde, L. & A. & Classification of hydration in clinical conditions: indirect and direct approaches using bioimpedance. *Nutrients* **11**, 809 (2019).
7. Khalil, S., Mohktar, M. & Ibrahim, F. The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases. *Sensors* **14**, 10895–10928 (2014).
8. Meroni, D., Bovio, D., Frisoli, P. A. & Aliverti, A. Measurement of electrical impedance in different ex-vivo tissues. in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2311–2314 (IEEE, Orlando, FL, USA, 2016). (2016). https://doi.org/10.1109/EMBC.2016.7591192

9. Toso, S. et al. Altered tissue electric properties in lung cancer patients as detected by bioelectric impedance vector analysis. **5**, (2000).
10. Baarends, E. M., Van Marken Lichtenbelt, W. D., Wouters, E. F. M. & Schols, A. M. W. J. Body-water compartments measured by bio-electrical impedance spectroscopy in patients with chronic obstructive pulmonary disease. *Clin. Nutr.* **17**, 15–22 (1998).
11. Baghbani, R., Moradi, M. H. & Shadmehr, M. B. Momayez Sanat, Z. A new Bio-Impedance forceps sensor for measuring electrical conductivity of the biological tissues. *IEEE Sens. J.* **19**, 11721–11731 (2019).
12. Baghbani, R., Shadmehr, M. B., Ashoorirad, M., Molaeezadeh, S. F. & Moradi, M. H. Bioimpedance spectroscopy measurement and classification of lung tissue to identify pulmonary nodules. *IEEE Trans. Instrum. Meas.* **70**, 1–7 (2021).
13. Sanchez, B. et al. In vivo electrical bioimpedance characterization of human lung tissue during the bronchoscopy procedure. A feasibility study. *Med. Eng. Phys.* **35**, 949–957 (2013).
14. Company-Se, G. et al. Minimally invasive lung tissue differentiation using electrical impedance spectroscopy: A comparison of the 3- and 4-Electrode methods. *IEEE Access.* **10**, 7354–7367 (2022).
15. Company-Se, G. et al. Effect of calibration for tissue differentiation between healthy and neoplasm lung using minimally invasive electrical impedance spectroscopy. *IEEE Access.* **10**, 103150–103163 (2022).
16. Company-Se, G. et al. Differentiation using Minimally-Invasive bioimpedance measurements of healthy and pathological lung tissue through bronchoscopy. *Front. Med.* https://doi.org/10.3389/fmed.2023.1108237 (2023).
17. May, M. Eight ways machine learning is assisting medicine. *Nat. Med.* **27**, 2–3 (2021).
18. Hosni, M., Carrillo-de-Gea, J. M., Idri, A., Fernandez-Aleman, J. L. & Garcia-Berna, J. A. Using ensemble classification methods in lung cancer disease. in. *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 1367–1370 (IEEE, Berlin, Germany, 2019). (2019). https://doi.org/10.1109/EMBC.2019.8857435
19. Bharati, S., Podder, P. & Mondal, M. R. H. Hybrid deep learning for detecting lung diseases from X-ray images. *Inf. Med. Unlocked.* **20**, 100391 (2020).
20. Tekerek, A. & Al-Rawe, I. A. M. A novel approach for prediction of lung disease using chest X-ray images based on densenet and MobileNet. *Wirel. Pers. Commun.* https://doi.org/10.1007/s11277-023-10489-y (2023).
21. Jasmine, P. et al. Lung Diseases Detection Using Various Deep Learning Algorithms. *J. Healthc. Eng.* 1–13 (2023). (2023).
22. Sreejith, S., Nehemiah, H. K. & Kannan, A. Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. *Comput. Biol. Med.* **126**, 103991 (2020).
23. Al-Zaiti, S. S. et al. A clinician's guide to Understanding and critically appraising machine learning studies: a checklist for ruling out bias using standard tools in machine learning (ROBUST-ML). *Eur. Heart J. - Digit. Health.* **3**, 125–140 (2022).
24. Palmieri, F. et al. Machine learning allows robust classification of visceral fat in women with obesity using common laboratory metrics. *Sci. Rep.* **14**, 17263 (2024).
25. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
26. Richardson, E. et al. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns* **5**, 100994 (2024).
27. Weinberger, E., Cockrill, S. A., Mandel, J. & B. & *Principles of Pulmonary Medicine* (Elsevier, 2019).
28. Hammer, D. & McPhee, J. G. S. *Pathophysiology of Disease, an Introduction To Clinical Medicine*. (McGraw-Hill Education) (2019).
29. Hassan, H. et al. Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. *Comput. Biol. Med.* **141**, 105123 (2022).
30. Müller, C., Guido, S. & A. & *Introduction To Machine Learning with Python, A Guide for Data Scientists* (O'Reilly Media,, 2017). United States of America.
31. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer New York Inc., 2001).
32. Chang, R. F., Wu, W. J., Moon, W. K., Chou, Y. H. & Chen, D. R. Support vector machines for diagnosis of breast tumors on US images. *Acad. Radiol.* **10**, 189–197 (2003).
33. Dreiseitl, S. et al. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *J. Biomed. Inf.* **34**, 28–36 (2001).
34. Gholamzadeh, M., Abtahi, H. & Safdari, R. Comparison of different machine learning algorithms to classify patients suspected of having sepsis infection in the intensive care unit. *Inf. Med. Unlocked.* **38**, 101236 (2023).
35. Hicks, S. A. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **12**, 5979 (2022).

Acknowledgements

To the patients without whom this study would not have been possible. In addition, Marta Navarro Colom, Laura Romero Roca, and Margarita Castro Jiménez from the Interventional Pulmonology Unit, Respiratory Medicine Department, Hospital de la Santa Creu i Sant Pau for the invaluable support.

Author contributions

GC: designed the experiments, performed the experiments, performed the data processing, ana-lyzed the data, drafted the manuscript, prepared the tables and figures, revised and approved the final version of the manuscript. VP: designed the experiments, performed the experiments, revised and approved the final version of the manuscript. AR: designed the experiments, performed the experiments, revised and approved the final version of the manuscript. PR: designed the experiments, revised and approved the final version of the manuscript. JR: designed the experiments, revised and approved the final version of the manuscript. RB: designed the experiments, revised and approved the final version of the manuscript. LN: designed the experiments, performed the data processing, analyzed the data, drafted the manuscript, prepared the tables and figures, revised and approved the final version of the manu-script.

Funding

This work was supported by the Spanish Ministry of Science and Innovation (PID2021-128602OB-C21) and supported by the Secretariat of Universities and Research of the Generalitat de Catalunya and the European Social Fund.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.N.

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.