Unsupervised learning techniques for differentiating healthy and diseased individuals with suspected respiratory diseases.

39603547

MSc Data Science

A dissertation submitted for the degree of
Master of Science in Data Science

Supervised by *Dr, James Stovold*

School of Computing and Communications
Lancaster University Leipzig

August, 2025

Declaration

I declare that the work presented in this dissertation is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. Estimated word count is: **10762**

Name: **39603547**
Date: **August, 2025**

Unsupervised learning techniques for differentiating healthy and diseased individuals with suspected respiratory diseases.

39603547, MSc Data Science.
School of Computing and Communications, Lancaster University
A dissertation submitted for the degree of *Master of Science* in Data Science.
August, 2025

Abstract

Spirometry is a medical technique that provides a comprehensive diagnosis in the evaluation of respiratory disease. Nowadays, clinical domain experts use spirometric sequences to extract the main lung function indexes and provide a direct diagnostic to patients under suspicion of respiratory disorders. However, these parameters are not always representative of the patient real situation, as they rely on other demographic and secondary factors. Several research studies have implemented supervised and unsupervised learning techniques, with the aim of creating models that effectively capture hidden insights from spirograms and permit to derive a clear and accurate clinical diagnosis in patients with restrictive and obstructive respiratory diseases. The current project ambitions to provide a hybrid approach that combines deep-learning based methods and unsupervised learning techniques to produce an effective clinical outcome with respect to patient status. Continuous spirometric data gathered by NHANES institution is utilized to perform this study. The proposed hybrid models extract relevant patterns from volume time series through autoencoders and reservoir computing and apply clustering methods on these compressed representations to distinguish between diseased and healthy patients. Models applied to raw spirometric recordings achieve clinically significant separation between healthy and diseased groups for a subset of 20 subjects.

Acknowledgements

I'd like to thank my academic supervisor, Dr. James Stovold for giving me the freedom to ajust this project to my own interests and for the support throughout the whole process. I would also like to thank my colleagues and teachers for everything that they have taught me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Respiratory disorders have been affecting humanity for decades. Spirometry has become one of the main tools for assessing the state of lung function in clinical practice. The diagnosis of respiratory diseases has evolved from manual interpretation of lung function parameters to the standardization of reference methods (Ong-Salvador, Laveneziana, and Jongh, 2024). Some of them have even been spread throughout the world and have a wide use, such as the multi-ethnic, age-spanning predictive equations, established by the Global Lung Function Initiative (GLI) (Quanjer et al., 2012).

The technological evolution in medical instrumental with the digitization of spirometers contributes to the application of data science-based approaches. These alternatives involve supervised learning techniques, introducing algorithms such as decision trees (DTs) or support vector machines (SVM), that serve as quality control and recognition of patterns (Giri et al., 2021).

Recently, deep learning techniques have been developed to model spirometry volume-time signals. These solutions are used for several different purposes, such as genomics, quality or acceptability assurance (Yimin Wang et al., 2022), and clinical decision making (Mac, Xu, Joyce K. Y. Wu, et al., 2022b). In the last one, most of the research has been conducted to extract relevant information from spirometry time series. This information is processed to feed downstream tasks, such as classification or clustering (Yimin Wang et al., 2022). This project aims to implement unsupervised learning techniques to assess patient status with respect to respiratory diseases. In this context, the latent space of autoencoders and the states of reservoir computing (RC) become powerful alternatives to traditional methods to retain the temporal dependencies and dynamics of spirometic sequences.
NHANES institution collects continuous data measurements, such as spirometric recordings, that can be used for research purposes (Centers for Disease Control and Prevention, 2024). These databases, compiled over several years, represent the basis on which this study is

developed. The first stage of the project provides a clinical overview of the spirometry test technique (Section 2.1) and the most relevant restrictive and obstructive respiratory diseases (Section 2.2). Then a literature review of related work is performed around the application of supervised (Section 3.1) and unsupervised learning (Section 3.2) algorithms applied to spirometric data.

The current work describes the structure and motivations of the NHANES institution and the spirometry tests databases it gathers (Section 4.1). All variables within these databases are carefully explored in Section 4.2, where some relevant correlations and associations between features are highlighted. Spirometric sequences are filtered and pre-processed on the basis of the common requirements established in the previous EDA (Section 4.3). These recordings are formatted and fed to the models detailed in Section 4.4. The final results are presented and discussed in Chapters 5 and 6 respectively. Finally, the main conclusions and future lines regarding the project are summarized in Chapter 7.

Chapter 2

Background

2.1 Description of the Spirometry Procedure

Under suspicion of respiratory disease, spirometry is a procedure that provides a comprehensive diagnostic in the evaluation of patients of all ages (Koegelenberg, F. Swart, and Irusen, 2012). This technique not only delivers a categorization of the disease, but also forecasts disease progression and tracks previous interventions. Prior to spirometry examinations, incorrect standards or operations significantly reduce the reliability of test results (Graham et al., 2019).

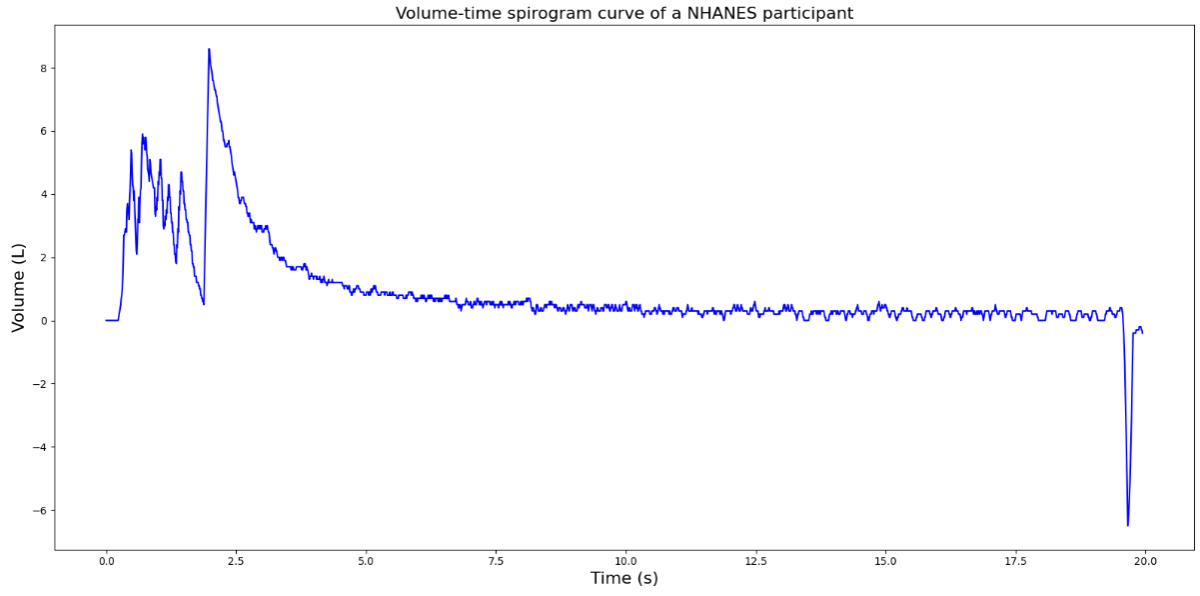

Figure 2.1: Volume-time spirogram curve of a NHANES participant.

Spirometry tests measure the volume and airflow within the lungs as a function of the physical properties of the respiratory tract, such as the lung parenchyma or the strength of the respiratory muscle. Modern spirometers allow for real-time digital representation of respiratory cycle maneuvers, which are displayed in the format of volume-time and flow-volume curves. These biomedical signals receive the name spirograms (Maree et al., 2022). Figure 2.1 shows a real example of a volume-time spirogram test for an individual from the NHANES datasets (Section 4.1).

2.1.1 Fundamental Spirometry Measurements

Spirometry tests provide a set of metrics that contribute to understanding the state of the patient's airways. Within these measurements, there is a distinction between the main parameters of lung function and complementary metrics (Maree et al., 2022). Subsequently, a list of the fundamental measurements that fall between these two groups is presented.

- **Main parameters of lung function**
  - **Forced Vital Capacity** (FVC) represents the maximum volume value, expressed in liters (L), that can be achieved in an air exhalation coming from a position of inspiration, requiring the greatest physical effort of the patient (Ponce, Sankari, and Sharma, 2023).
  - **Forced Expiratory Volume in the first second** (FEV) corresponds to the total exhalation of gas, in L, only taking into account the first second of the FVC maneuver (David, Goldin, and Edwards, 2024).
  - **Forced Expiratory Ratio** (FER) expresses the rate between FEV and FVC. In other words, the proportion of air exhaled from the FVC that is ejected during the first second of the forced maneuver (Ponce, Sankari, and Sharma, 2023).
  - **Peak Expiratory Flow** (PEF) is the maximum flow value reached in the course of the FVC maneuver (Haynes, 2018).
- **Complementary metrics of lung function**
  - **Vital Capacity** (VC) refers to the amount of air volume that can be inhaled or exhaled during a regular breathing maneuver (Ponce, Sankari, and Sharma, 2023).
  - **Forced Expiratory Flow** (FEF) can be estimated at several points where a certain proportion of air has been exhaled. It is used to understand in which phase of the FVC the flow varies the most (Maree et al., 2022).
  - **Total Lung Capacity** (TLC) corresponds to the total volume of air that the lungs can hold after maximal inhalation.

The main parameters and complementary variables of lung function represent a valuable clinical tool for clinical experts to characterize respiratory tract functioning and identify abnormalities within respiratory biosignals. Calculating these metrics arises as intangible information that improves diagnosis and treatment (Hoesterey et al., 2019).

2.1.2 Essential Requirements for Spirometry Testing

Spirometry reports should contain a minimum of patient information, including demographic data such as name, sex, age, ethnic group, etc., in addition to the spirogram curves. The test must be performed under ambient conditions, considering a posterior correction of the recordings applying the Body Temperature and Pressure, Saturated (BTPS) factor (M R Miller et al., 2005a). The spirometer must be calibrated at regular time intervals and properly disinfected between every patient trial. Both the administration of the medication and the position of the patient during the examination affect the outcome of the test. The individual subjected to the test must avoid smoking and intense exercise at least one hour prior. Furthermore, the collaboration and attitude of the subject is an important factor in the clinical reliability of the test (Maree et al., 2022).

A single individual must perform at least three acceptable maneuvers. An acceptable maneuver is defined by a full inspiration and expiration cycle that is followed by a plateau in the spirogram series. Moreover, the signal is expected to be exempt from artifacts and fluctuations and to devoid of hesitation at the beginning of the recording (Masekela et al., 2013). Finally, repeatability of the test ensures an effective patient performance during examination (M. R. Miller et al., 2005b).

2.1.3 Clinical Interpretation of Spirometric Curves

Clinical experts usually use derived spirometric calculations (described in Section 2.1.1) to determine whether there is an abnormality in the functioning of the respiratory tract. In pulmonary medicine, the Lower Limit of Normal (LLN) is defined as the reference constant, where fewer 5 % of the population FER values fall below that value. The American Thoracic Society fixes this value to 0.7, which means that patients who show FER lower than 0.7 present obstructive patterns (Stanojevic, Kaminsky, Martin R. Miller, et al., 2022). Otherwise, a restrictive pattern is characterized by the combination of a reduced FVC and elevated FER. This clinical decision is confirmed through the TLC exploration. Eventually, the mixed effect, restrictive and obstructive behavior, appears when both FER and FVC present decreased values (Barreiro and Perillo, 2004). Table 2.1 summarizes the association of the pattern nature with the main parameters of lung function.

Table 2.1: Obstructive and restrictive patterns comparative analysis (Al-Ashkar, Mehra, and Mazzone, 2003).

Measurement	**Obstructive Pattern**	**Restrictive Pattern**
Forced Vital Capacity (FVC)	Decreased or normal	Decreased
Forced Expiratory Volume in 1 second (FEV_1)	Decreased	Decreased or normal
FEV_1/FVC ratio	Decreased	Normal
Total Lung Capacity (TLC)	Normal or increased	Decreased

2.2 Obstructive and Restrictive Pulmonary Disorders Overview

In a clinical context, several respiratory conditions can be classified as obstructive or restrictive diseases, exploring physiopathological differences between them. As mentioned in Section 2.1, spirometers translate physiological differences into biomedical signals that can be analyzed. Nevertheless, some respiratory diseases share underlying characteristics that make the diagnosis unclear at first glance. Dyspnea is a multi-factorial breathing discomfort that varies in intensity (American Thoracic Society, 1999). This condition generates common breathing mechanisms between obstructive and restrictive respiratory disorders. An example of this is that patients require to exert more effort to achieve air movement through the respiratory tract or the mechanics of the chest wall are forcefully altered(Scano, Innocenti-Bruni, and Stendardi, 2010).

2.2.1 Obstructive Lung Disorders

Obstructive respiratory diseases affect the lungs and airways structurally over time. In the case of muco-obstructive conditions, the mucus layers migrate to the trachea and the accumulation of mucus attached to the walls of the airways produces sputum or phlegm (Singh et al., 2023). The other obstructive respiratory disorders involve structural loss, inflammation, or airway collapse (Cukic et al., 2013). The main obstructive respiratory conditions are briefly described below.

- **Chronic Obstructive Pulmonary Disease** (COPD) is characterized by massive mucus secretion that causes chronic obstruction of the airways (Singh et al., 2023).
- **Cystic Fibrosis** (CF) is a genetic disease that generates dehydrated mucus production that leads to recurrent infections and obstruction of the airways (Singh et al., 2023).
- **Asthma** is explained by airway obstruction caused by bronchial hyperreactivity and smooth muscle contraction (Cukic et al., 2013).

- **Primary Ciliary Diakinesia** (PCD) consists of a ciliary dysfunction that promotes unintentional retention of mucus in the respiratory tract (Singh et al., 2023).

2.2.2 Restrictive Lung Disorders

Restrictive respiratory diseases are characterized by the induction of reduced lung volumes, which is observable through TLC values and the reduction in lung expansion capacity, as expressed in Table 2.1. Affectations such as interstitial lung disease (ILD) or chest wall damage are some of the repercussions of these types of disease (Martinez-Pitre, Sabbula, and Cascella, 2025). Some kind of restrictive respiratory disorders are commented on in the following list.

- **Idiopathic Pulmonary Fibrosis** (IPF) remains a mystery to researchers and leads to lung tissue injury and inadequate gas exchange (Martinez-Pitre, Sabbula, and Cascella, 2025).
- **Sarcoidosis** is a multi-organ formation of granulomas that leads to restrictive lung behavior (Martinez-Pitre, Sabbula, and Cascella, 2025).
- **Pneumoconiosis** is the result of the inhalation of different types of dust that originates lung injuries and restrictive characteristics (Cohen, Patel, and Green, 2008).
- **Obesity Hypoventilation Syndrome** (OHS) is a condition derived from obesity, which avoids correct breathing and promotes restrictive patterns (Backman et al., 2016).

Chapter 3

Literature review

3.1 Supervised Learning for Respiratory Disease

3.1.1 Pulmonary Function Parameters into Diagnostic Models

Classical approaches to detect respiratory diseases such as COPD consist of evaluating lug function parameters from spirometric recordings (Barnes and Fromer, 2011). Not all spirograms provide a direct calculation of variables of lung function. Approximately 50 % of the trials are considered inconsistent (Schermer et al., 2003). Different approaches such as regression curves for parameter estimation, artificial neural networks (ANN) applied to patient demographic and personal data, or multi-output support vector regression (SVR) models (Zhou et al., 2022) have been developed to improve parameter estimation.

Forced oscillometry (FOT) presents a non-invasive solution (Bickel et al., 2014) that provides regression curves to estimate parameters such as FVC and FEV (described in Section 2.1.1) (Yamamoto et al., 2017). FOT indices, combined with the previous knowledge of the patients, generate regression curves that assess the main parameters of lung function. Multivariate linear regression delivers a good correlation of measured values within several diseases and a diagnostic accuracy of 88 % (Miyoshi et al., 2020). Despite the accurate results, the estimated curves do not match the unseen data. The idea of using pulmonary function parameters contrasts with the approach of this project, which aims to provide a direct clinical outcome from spirometric data.

AAN is introduced to derive the area under the expiratory flow-volume curve (AEX) from the basic information of the patients (age, sex, etc.). The model achieves a R^2 of about 0.8 in two different cohorts (Ioachimescu, Stoller, and Garcia-Rio, 2020). Although AEX appears to be an excellent indicator of abnormalities in lung function (Vermaak, Bunn, and Kock, 1979), this method still requires a previous calculation of lung function parameters,

increasing the chances of missing early detection of the disease.

Multi-output SVR has been applied to a set of demographic characteristics and inflammatory parameters to determine FEV and FVC (described in Section 2.1.1) (Vapnik, 2010). Severe, non-severe, and mixed SVR models are constructed. Figures 3.1 and 3.2 provide a graphical representation of prediction performance in the testing sets for the three different models. The non-severe and severe models outperform the mixed model. The non-severe model stands out from the other models with an acceptance rate of 90 %. Despite that, the overall values of R^2 remain lower than 0.4 (Chen et al., 2020). In general, creating two different models in terms of severity does not capture all the patterns in a global analysis of the complete signal.

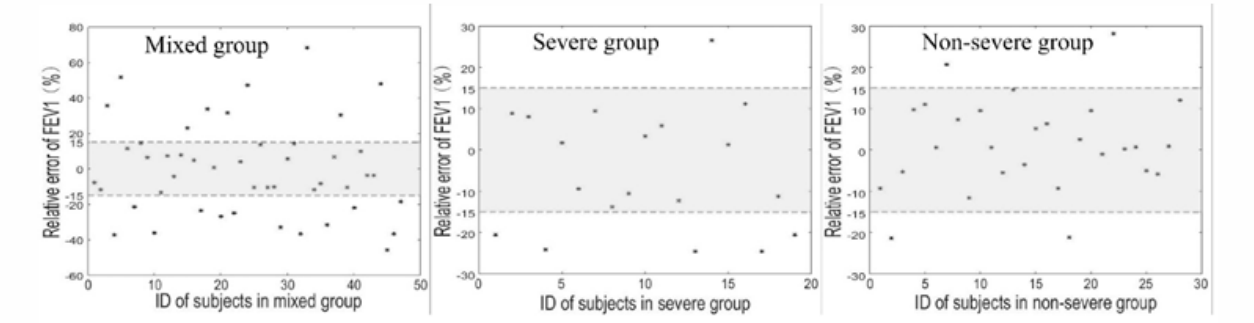

Figure 3.1: FEV prediction performance (Chen et al., 2020).

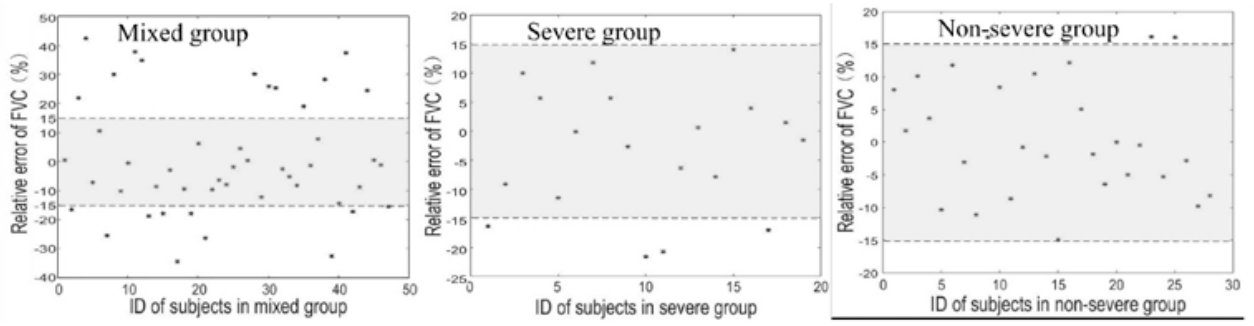

Figure 3.2: FVC prediction performance (Chen et al., 2020).

Random Forrest (RF) appears as another alternative machine learning model, using lung function parameters (FEV and FVC) to predict respiratory tract health. Kristensen et al. (2023) apply RF models to approximately 23000 NHANES (Section 4.1) participants. The

correlations between ground-truth FEV and FVC and the estimated values are found to be $R^2 = 0.92$ and $R^2 = 0.95$, respectively (Kristensen et al., 2023).

3.1.2 Incorporating Spirometric Data in Computational Diagnosis

Traditional methods involve passing through the parameters of lung function to reach a clinical outcome. These alternatives are not accurate enough for people in different age groups, hindering solutions such as personalized treatment (Bhatt et al., 2023). DeepSpiro is a model that aims to predict and prevent COPD. DeepSpiro implements deep learning to volume and flow time series from spirometry examinations. Figure 3.3 provides a clear summary of the main components of the architecture of the model. Labeled data from spirometry recordings containing COPD cases from the UK Biobank are introduced to the model (Mei et al., 2025). The results on unseen data provide an area under the curve (AUC) value of 0.83, describing an overall strong discriminative power, while sensitivity gives a value around 0.71, indicating a margin of improvement in detecting positive cases.

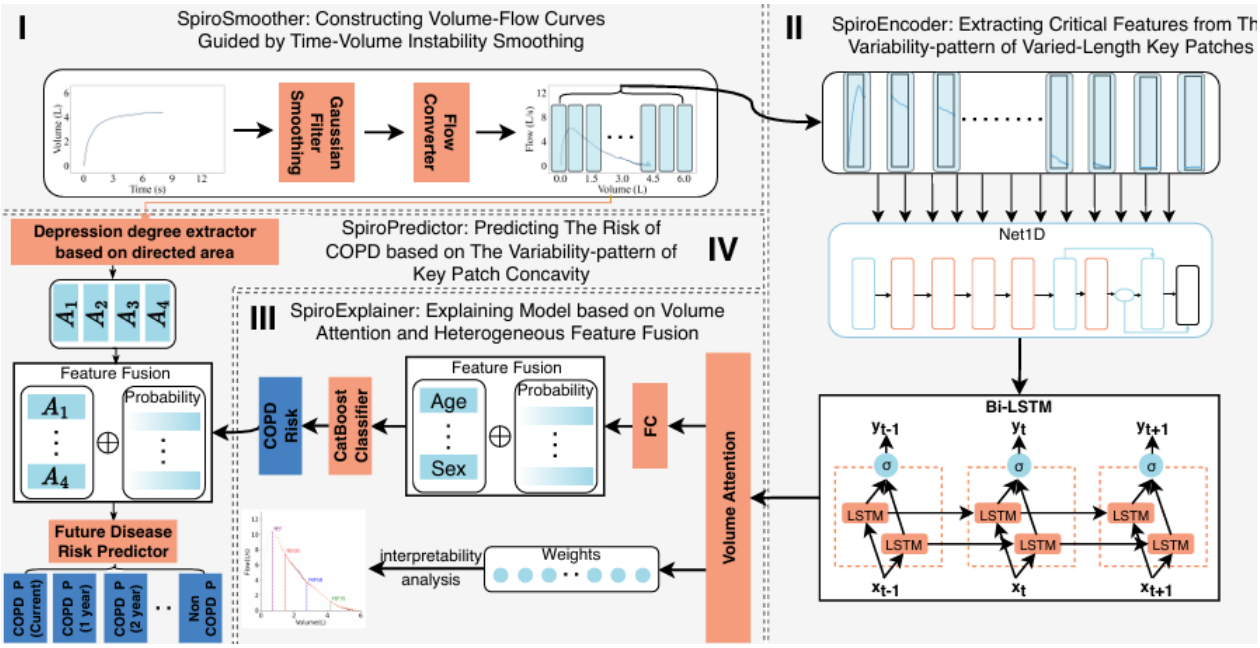

Figure 3.3: DeepSpiro framework overview (Mei et al., 2025).

Different machine learning algorithms are adapted to simulated spirometric data. Models such as andom Forrest (RF) and Support Vector Machine (SVM) provide an accuracy of 99 % (Di Dio et al., 2021) in these data. In general, all different models have an excellent fit to the data, which can be observed in 3.4. Perhaps, these results clearly reflect overfitting

without exposing the models to unseen data from real patients. The current findings are of poor clinical significance.

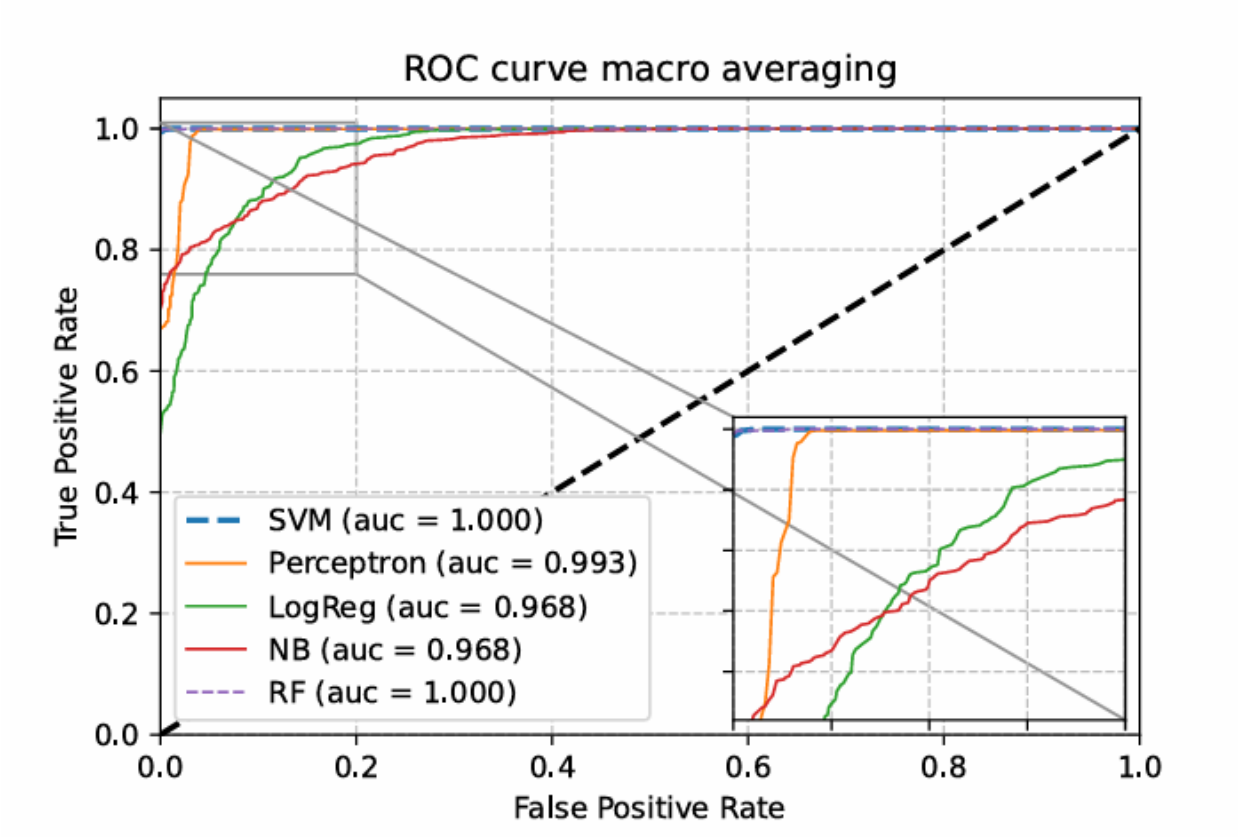

Figure 3.4: ROC curves of machine learning models applied to spirometric simulated data. (Di Dio et al., 2021).

Time-series transformers emerge as a possible alternative to DeepSpiro's solution, accessing spirometric recordings of COPD patients from the UK Biobank. The results provided by this method are slightly better, as the AUC delivers a value of 0.83, which is an excellent fit (Gadgil, Galanter, and Negahdar, 2024). The main concern with this method is the high complexity of the method and the need for demographic data to improve the predictive accuracy.

Moreno Mendez et al. (2024) implement gradient boost on a labeled COPD smokers dataset, aged between 30 and 80. Obstructive patters are captured with an accuracy of 95 % and a sensitivity of 93 %. However, a population bias can be introduced, as the work only considers

smokers and avoids other population groups (Moreno Mendez et al., 2024).

Mac, Xu, Joyce K Y Wu, et al. (2022a) have built a multi-perceptron model fed with spirometric and pletismographic recordings and biometrics from 748 patients. The study compares the results of the test after training the neural networks with mixed data. After refining the model, it shows an accuracy of about 95%, which improves the DT models and the diagnostics of clinical experts (Mac, Xu, Joyce K Y Wu, et al., 2022a). Unfortunately, the model requires the combination of several data resources.

3.2 Unsupervised Learning for Respiratory Disease

Several studies explore the combination of unsupervised learning and demographic and spirometric data to predict respiratory diseases. Yuan et al. (2022) have recovered a combination of computed tomographies (CT), demographic, and spirometric variables from up to 916 different phenotypes. The study reveals that about 65 % of the data variability is explained by CT scans. These results illustrate that unsupervised learning methods might be more efficient to apply to CT scan data than to spirometric time series (Yuan et al., 2022). However, spirometry tests are more accessible, cost-effective and safe than medical imaging tools such as CT (National Guideline Alliance (UK), 2017).

A research published in the Indian Journal of Science and Technology in 2016 attempts to combine K-Means clustering and DT algorithms. The resulting clusters are expected to represent similar lung functions in each group. Then, both cluster labels and lung function characteristics are introduced in the DT (Karuppiah and Suseendran, 2016). There is no clear performance recovered from this article, as the only detailed steps correspond to the methodology. No validation of the algorithm is concluded; thus, this hybrid approach is not directly implemented.

Scrucca et al. (2016) applied finite Gaussian mixture modeling to lung function parameters to generate clusters of patients exhibiting similar respiratory behaviors. Data collected from the Rotterdam study between 2002 and 2016 are used for this purpose. Multiple groups are obtained with different lung function trajectories. 43 out of 100 patients show persistent normal behavior, while close to 30 % reflect a high FEV (Bertels et al., 2024). This model does not provide direct results on patient status, but reveals clinical information on lung function from patients.

Thanks to data gathered from the Chance study, Augustin et al. (2018) conducted a research to meet different subgroups of hospital interns with common patterns of lung function through self-organizing maps (SOM). SOM groups patients with similar attributes of lung function.

After implementing SOM on the data, 7 groups correspond to varying airflow limitations are created. The ultimate objective of the study is not related to delivering a diagnostic to patients, but to improve clinical decisions of healthcare professionals.

All these research studies provide a set of points which clarify why implementing unsupervised learning methods to the raw spirometric sequences is convenient, rather than following traditional pathways:

- Taking the main parameters of lung function as input data for the detection of respiratory diseases is a valid approach, but it implies losing relevant information of hidden patterns within the complete spirometric data.
- Parameters such as FEV or FVC are not representative for patient groups that have different ages or ethnicities in some studies.
- Deploying models that use simulated or augmented data is not realistic and can be prone to overfit and fail in predictions on unseen data from patients coming from a completely different context.
- Supervised methods require previous trials in patients, which is a time, cost, and personnel expense that can be alleviated by properly implementing unsupervised learning techniques.
- Most of the models are applied to a combination of demographic, lung function parameters, and even medical imaging trials, which involve a large amount of data, with high production and personnel costs.
- Unsupervised learning techniques focus on providing disease-related information instead of an explicit clinical outcome, clarifying whether patients are healthy or not.

Chapter 4

Methodology

4.1 NHANES Dataset

This project is supported by the National Health and Nutrition Examination Survey (NHANES) institution. This center gathers measurements on the nutrition and health status of the population of the United States of America (USA), including health and laboratory trials for all ages. In NHANES, data collection is mediated by a randomized scientific protocol, in such a way that the entire USA population can be represented by recovered samples (Centers for Disease Control and Prevention, 2024). Some multidisciplinary projects, such as prediction of distributional profiles of physical activity (Matabuena, Ghosal, et al., 2023), distributional data analysis of accelerometer data (Matabuena and Petersen, 2021) or a construction of a dataset of cancer patients for epidemiology research (Moon and Mun, 2025) have benefited from the available data provided by NHANES.

Three NHANES datasets containing raw data curves from spirometry trials in patients between the ages of 6 and 79 were built in 2007-2008, 2009-2010, and 2011-2012 periods. These databases combine a healthy population with patients undergoing chronic obstructive pulmonary disease (COPD). All collected spirometric sequences were unedited and follow the American Thoracic Society (ATS) standards. The volume-time curves of the spirometry trials are compiled together with several clinical data related to the test (Tilert et al., 2013). Table 4.1 provides a brief description of all features contained in the NHANES spirometry datasets, as well as the data type for each variable.

Table 4.1: Description of variables in the NHANES Spirometry Datasets

Variable	**Data Type**	**Name**	**Description**
SEQN	float	Respondent Sequence Number	Unique identifier for each survey participant.
SPATTYPE	object	Type of Test	Indicates whether the test is pre- or post-bronchodilator.
SPAMANU	float	Spirometry Curve Number	Curve number in the spirometry test.
SPAFACT	float	BPTS Correction Factor	Correction factor applied to the raw data.
SPAPOS	string	Testing Position	Participant's position during testing (standing or sitting).
SPAPLAT	string	Plateau Achieved	Whether a plateau was achieved in the individual raw curve.
SPAACC	string	Acceptable Maneuver	Indicates if the maneuver is acceptable.
SPAQEFF	string	Examinee Effort	Assesses the effort of the examinee during the test.
SPXPTS	float	Number of Data Points	Number of data points in the individual raw spirometry curve.
SPXRAW	string	Raw Spirometry Data	Comma-delimited raw spirometry curve data.

Patient information is stored in relation to a single identifier (SEQN in Table 4.1) and most patients are subjected to more than one spirometry test. In Table 4.1, SPATTYPE refers to the treatment applied to the patient before or after the test. A pre-bronchodilator can be administered to the participants before initiating the respiratory cycle. This medication is mainly applied to observe whether the medication improves airflow. If an improvement is observed, the patients may have a restrictive disease (Mannino, Diaz-Guzman, and Buist, 2011). All spirometry tests are recorded under ambient conditions, which is not representative of the physiological scenario inside the lungs. A correction in the original signals must be applied to adjust the recordings to body temperature, pressure and water vapor air saturation conditions (BTPS) (Graham et al., 2019).

Within all factors mentioned in Table 4.1, the position of the test also affects the final result. The standing position is known to allow for better lung expansion, while the sitting position ensures patient safety, especially for elderly patients with serious health problems (Lalloo, Becklake, and Goldsmith, 1991). In Table 4.1, SPAPLAT informs whether the plateau is reached in the test or not. The plateau indicates that the subject has fully exhaled all the air contained in the lungs, and thus the FVC is adequately assessed (M R Miller et al., 2005a). In addition, to ensure that the maneuver is performed correctly, it is critical that the subject achieves the maximum effort during the test. Otherwise, lung function can be underestimated, leading to an incorrect diagnosis (M R Miller et al., 2005a). Finally, the

test values are stored in a string that spreads out every measurement point by a comma. These databases do not include labels regarding the possibility of having individuals with restrictive or obstructive conditions.

4.2 Exploratory Data Analysis

The three versions of NHANES spirometric examinations (detailed in Section 4.1) are merged into a single dataset to have access to a larger source of tests. These datasets with common variables are simply concatenated, from the oldest to the most recent instances. This present work takes into account all the data provided from NHANES, without introducing any kind of bias.

4.2.1 Univariate Analysis

The total number of instances of the final dataset is 108939 recordings, with 10 variables recovered as shown in Table 4.1. The dataset reveals that, in some cases, more than one test is applied per participant, which means that some of the trials might be duplicated. The mean number of tests performed on every individual is 3.3 and at least 50 % of the samples have been tested three times, even if the test conditions were different between examinations (presence or absence of pre-bronchodilator). The maximum number of tests received by a single patient is 13. The BPTS correction factor presents low variability, with a standard deviation of about 0.01. Almost all tests are carried out under similar conditions and no technical errors in the data generation from the spirometers are found. The number of time steps recorded in every trial has a mean value of 857. However, some samples seem to be too short to contain a full expiratory cycle, with some sequences constituted of a single time step. Table 4.2 summarizes all the basic statistics on the variables in the dataset.

Table 4.2: Descriptive statistics on numerical variables of NHANES datasets

Statistic	**SPAMANU**	**SPAFACT**	**SPXPTS**
mean	3.31	1.08	857.00
std	1.92	0.01	411.50
min	1.00	1.04	1.00
25%	2.00	1.07	578.00
50%	3.20	1.08	848.00
75%	4.00	1.09	1097.00
max	13.00	1.15	2044.00

Figure 4.1 provides a clear overview of how the data is distributed and what are the extreme

values on the correction factor and the number of points per sequence. BPTS correction factor only varies within a range of values of (1.04 - 1.15). Hence, all sequences are measured under similar temperature and pressure conditions inside the spirometer. The distribution of the number of points per record suggests that there are a significant number of sequences which are too short to confirm that the entire expiratory cycle has been performed.

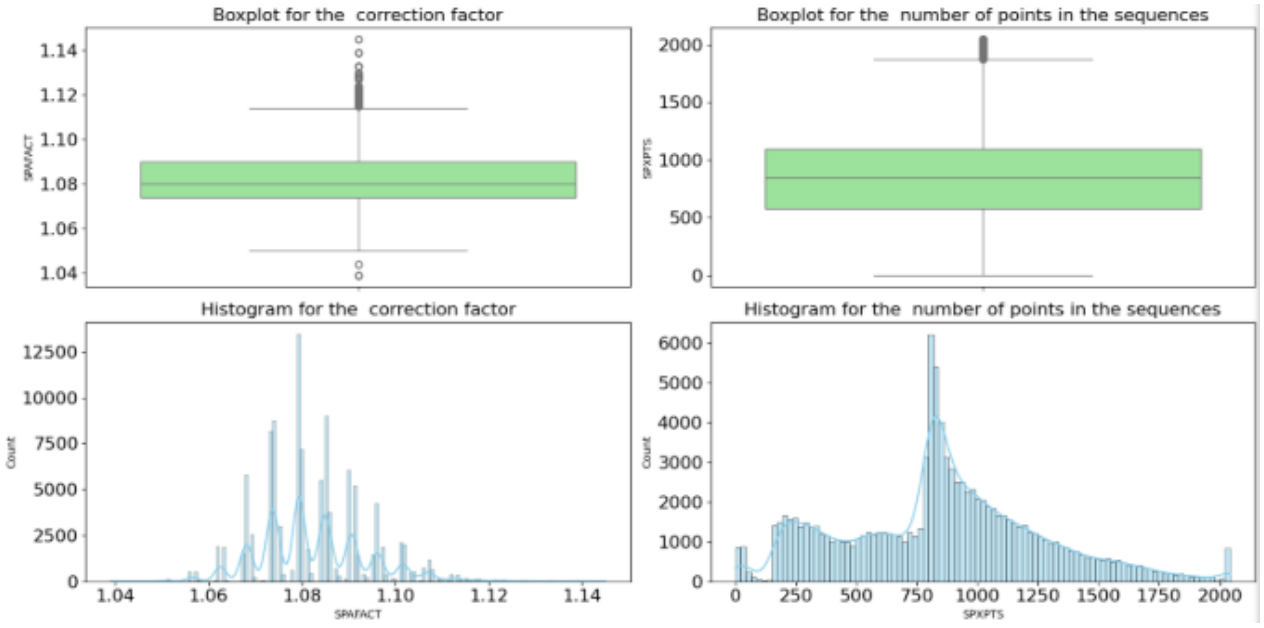

Figure 4.1: Boxplot and histogram of BTPS correction factor (SPAFACT in Table 4.1) and the number of time steps per recording (SPXPTS in Table 4.1).

The recovered data present imbalances between classes. Approximately 6 % of the trials are administered without a pre-bronchodilator, which means that in most cases, from the second measurement onward, specific medications are applied to improve the opening of patient's airways. In addition, close to 0.2 % of the tests are performed in the sitting position. In other words, there is an extremely small percentage of seriously ill patients going through the trials. Regarding the quality of the sequences, one in every four does not reach the plateau. Put differently, almost 25 % of the expiratory cycles are not fully completed.

Beyond that, there is a small proportion (around 7.5 %) of tests that are described as a non-acceptable maneuver. Figure 4.2 displays the effort of the participant, which is introduced as an ordinal categorical variable, in which "A" denotes an excellent effort, while a "D" reflects a poor performance. As shown in Figure 4.2, more than half of the examinees reach the best grade, while the second highest score achieved is a "C".

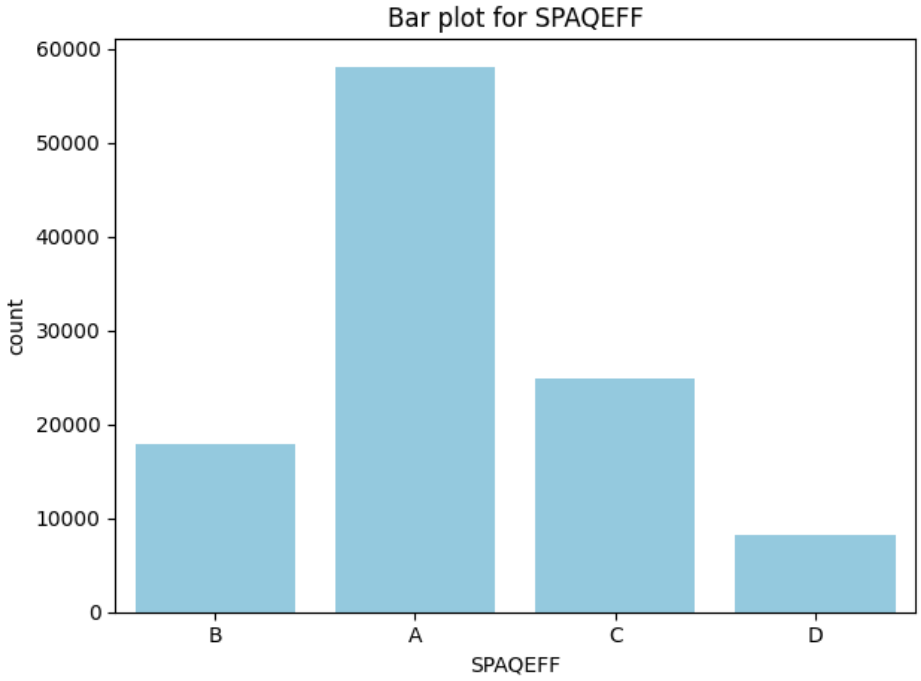

Figure 4.2: Barplot of the examinee effort.

4.2.2 Multivariate Analysis

By analyzing several variables at the same time, it is possible to understand how they are correlated with each other. To illustrate this, most sequences that do not reach the plateau are short in length. As shown in 4.3, most sequences without a plateau are placed in a range between 0 and 500 time steps, while time series reaching the plateau present a number of time points between 500 and 1500. There is a direct relationship between the number of points and the completion of the full expiratory cycle. In addition, a reduced number of time steps is usually related to a decreasing effort of the examinee. The test grades of "A" or "B" are more prevalent between 500 and 2000 time points, whereas low grades such as "C" or "D" appear in sequences with a decreasing number of points (4.3).

The acceptability status of the spirometry maneuver is correlated to the number of data points in the volume-time curves. Violin plots in Figure 4.4 highlight that the median value of points is higher for acceptable tests. In Figure 4.4, defective tests exhibit a more concentrated distribution of data instances at low time step values. In contrast, the shape of the violin plot for proper maneuvers is broader for higher values, extending around the value of 800 data points. As observed in boxplots of Figure 4.4, the same trend is reproduced for the number of curve points as a function of the medication administered. Examinations in which a pre-bronchodilator is administered to the patient are associated with shorter recordings

and vice versa.

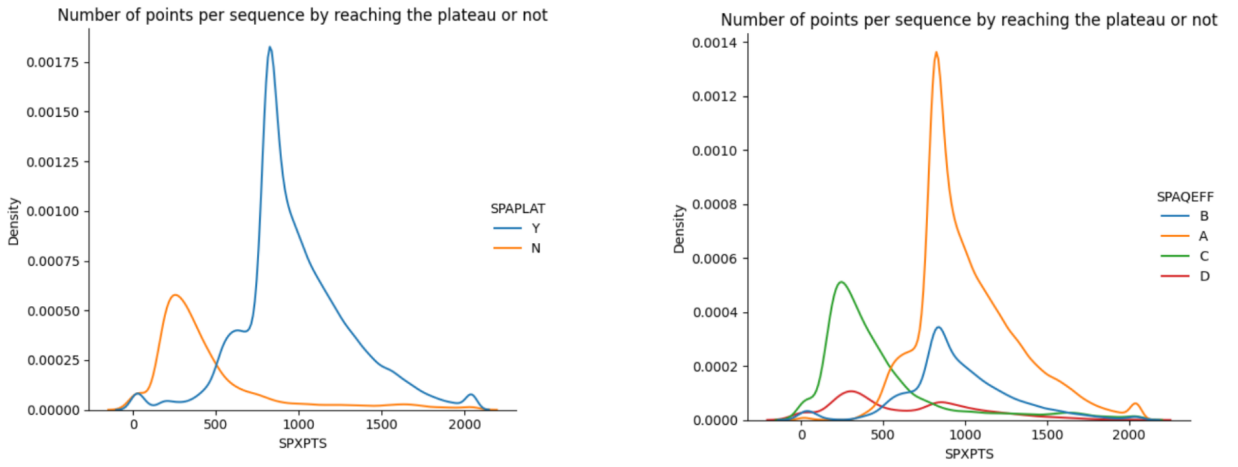

Figure 4.3: Density plots of the number of sequence points as a function of the the completion of the expiratory cycle and the examinee effort.

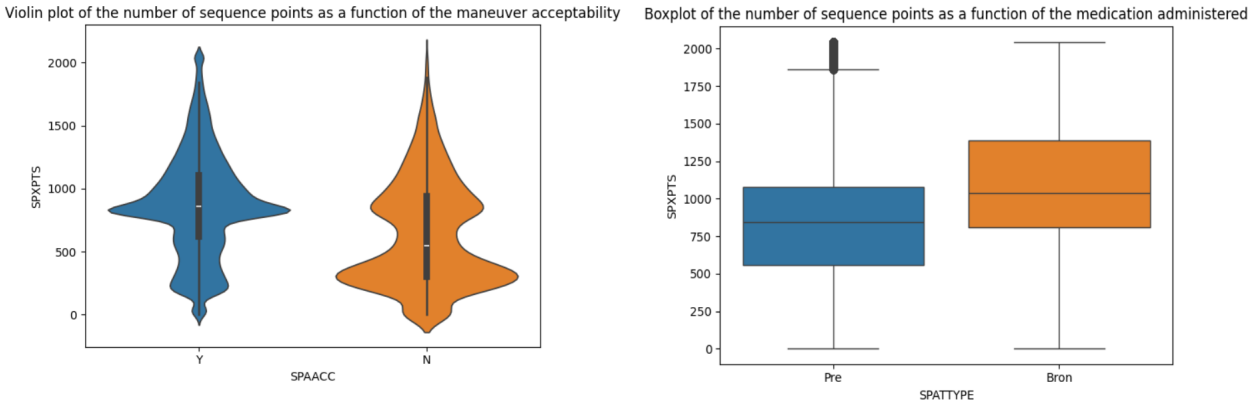

Figure 4.4: Representing the number of sequence points against the maneuver acceptability and the medication administration.

Pearson and Spearman correlation evidence that there is no correlation between the number of time steps recorded in a sequence and the BPTS correction factor. Pearson and Spearman correlation metrics are used in continuous variables to detect linear associations and non-linear or monotonic relationships between features (Azman et al., 2006). Equations 4.1 and

4.2 represent Pearson and Spearman's equations. In equation 4.1, x_i and y_i are the number of points per sequence and the correction factor, while $\bar{x}$ and $\bar{y}$ correspond to the mean values of both features. Equation 4.2 contains n, the number of paired observations in the data, and d_i as the difference between the ranks of the two variables for every observation. In this case, r_{xy} shows a value below 0.005 and ρ a value of about 0.0002, which indicates that there is no correlation between these two variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.1}$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{4.2}$$

The correlation between the categorical variables of the dataset is explored using Cramér's equation. In equation 4.3, χ^2 computes the test of independence for the contingency table on the combination of variables. χ^2 examines whether the distribution of one variable relies on the other or not. The rest of the parameters are the number of observations (n) and the number of columns (k) and rows (r) (Kim, 2017).

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1,\, r-1)}} \tag{4.3}$$

The achievement of the plateau on the tests is found to be highly correlated ($V \approx 0.9$) with the effort of the participant during the spirometry examination (Figure 4.5). Furthermore, there is a perfect relationship between the effort of the examinee and maneuver acceptability ($V \approx 1$). The results summarized in 4.5 demonstrate that the acceptability of the test is based on obtaining a high score in the effort category, which is simultaneously dependent on the achievement of the plateau. Beyond that, the remaining categorical variables in the dataset are uncorrelated.

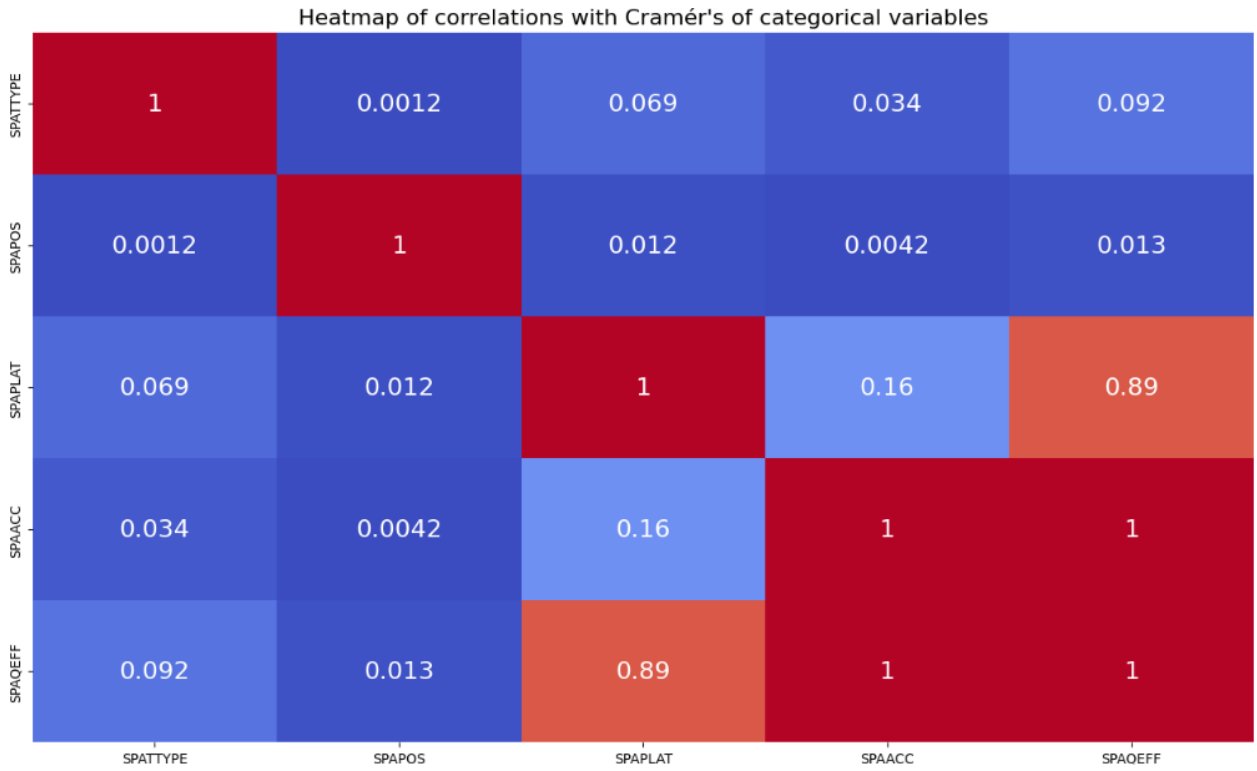

Figure 4.5: Heatmap of Cramér's correlations between NHANES categorical variables.

4.2.3 Findings and Insights from EDA

Based on exploratory data analysis, the key findings of the univariate and multivariate analysis are summarized below.

- Most patients are subjected to several spirometry tests under different medication and position conditions, and some sequences are too short to represent complete expiratory cycles.
- There are significant differences in the frequency of individuals evaluated in certain positions and medications. Most participants are given a pre-bronchodilator and perform the spirometry test in a standing position.
- The test validation is mediated by the sequences with a plateau. The acceptability of the spirometry maneuver and the effort of the examinee determine which are the valid sequences.

- Recordings associated with low effort scores or that do not reach the plateau correspond to short volume-time recordings.
- Longer data point sequences appear in acceptable maneuvers and for patients who have not been subjected to bronchodilator treatment prior to spirometry.
- The tests are validated if both a high score in the effort of the participant and the plateau are attained.

4.3 Data Preprocessing and Integration

After merging the three NHANES datasets, a total of 108939 observations are conserved. Neither missing column records nor null values are encountered. However, 138 duplicated rows are found within the dataset. These duplicates correspond to individuals whose spirometry recordings register a very low number of points. Duplicates are identified by filtering the dataset by patient ID, number of points, and volume-time recordings. Since these sequences are too short to be considered, the duplicates are deleted from the data, even though some are labeled as acceptable or present a good effort grade.

4.3.1 Volume Spirograms Conversion and Integration

The next step is to convert the strings that contain one-by-one the time-series values (SPXRAW in Table 4.1) into usable sequences. Therefore, string elements are converted to numerical values and introduced into lists. All sequences are encoded as lists of real numbers that actually represent the volume values. The sampling rate of the spirometric waveforms is set at 100 Hz. The sampling rate is the number of data points recovered per unit of time (Fallahtafti, Wurdeman, and Yentes, 2021). In this case, the sampling rate equation is used to calculate the period of the sequences, which is the difference of the time steps between two consecutive data points. In Equation 4.4, T represents the period, which is the inverse of the sampling rate (f_s). The spirometer records volume instances every 10 milliseconds (ms).

$$T = \frac{1}{f_s} \tag{4.4}$$

The representations of the volume-time curves reveal the presence of noise. Although artifacts in biological signals can be the result of biological variability itself, measurement error, or even environmental interference, spirometric sequences are affected by the sampling and digitization noise effect. This quantization noise is generated when a continuous signal is stored as discrete samples (Sadiya, Alhanai, and Ghassemi, 2021). Moving average filters

are selected to solve defects that appear in the spirometric signals. The implementation of these types of filters is easy; they reduce digitization and sampling noise, providing a smooth signal without the need for complex filter compositions (Pandey et al., 2021). The selection of the window size in the average moving filters conditions the amount of detail retained in the signal, as well as the conservation of the peaks and the global shape.

For the sake of the project, although residual noise slightly remains in the data, small window sizes are introduced to maintain temporal patterns in the volume-time recordings, which are critical to understanding whether individuals are healthy or not. Figure 4.6 illustrates a noisy representation of volume time series, in contrast to Figure 4.7, which shows a denoised signal from the same participant applying a window size of 20. The window size is the total number of points that, averaged together, aim to reduce the noise of the sequence.

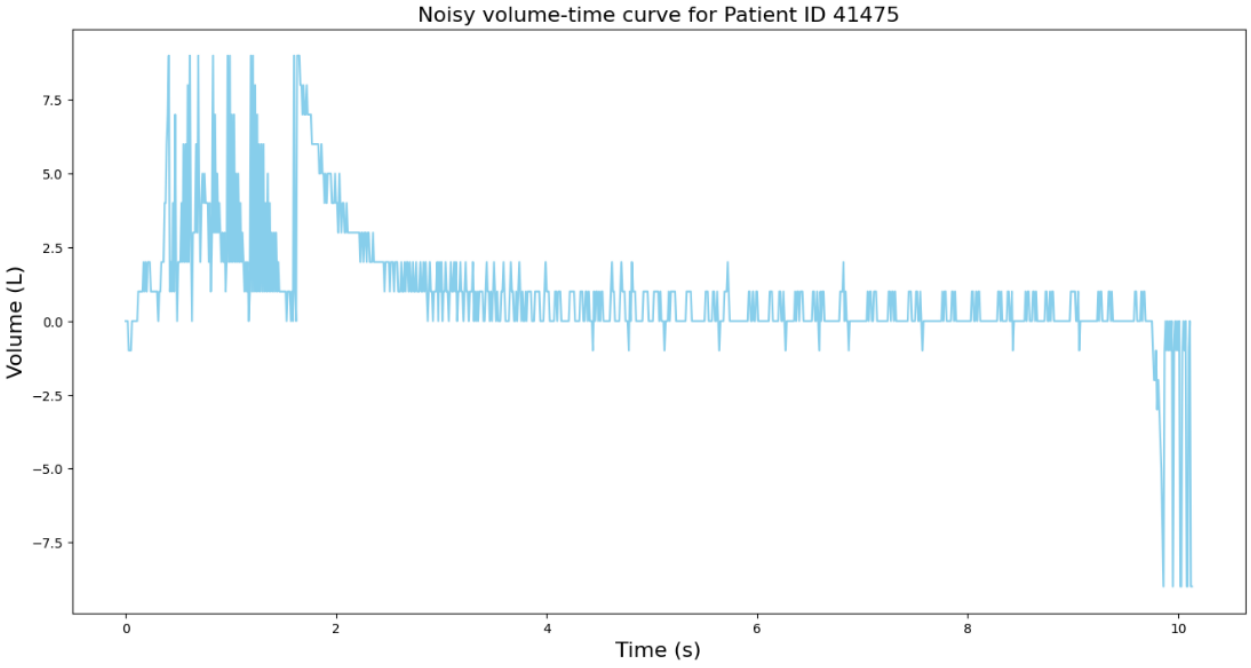

Figure 4.6: Noisy volume-time signal representation.

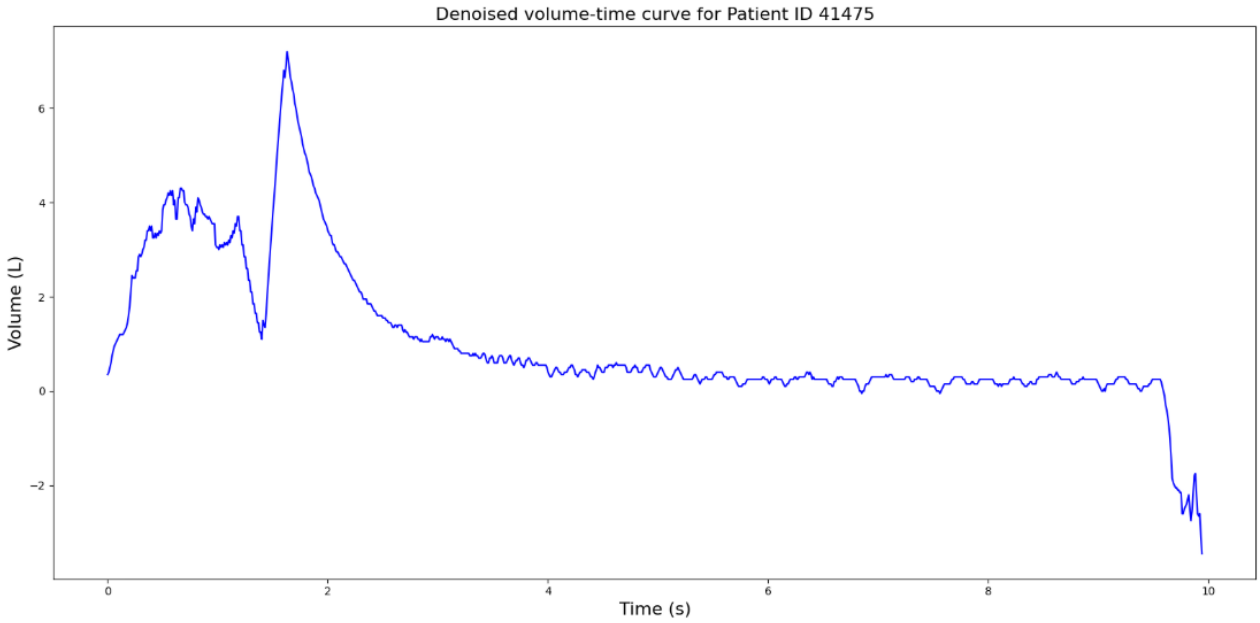

Figure 4.7: Denoised volume-time signal representation.

4.3.2 Retention of Quality Spirometric Curves

In Section 4.2, the EDA unveils which sample groups are worth preserving from the whole dataset. The cleaned dataset is then filtered to obtain, on one hand, the valid sequences in a list and, on the other hand, the patient ID's of these spirometry curves. Table 4.3 shows the conditions that must be met for the signal to be considered valid. In summary, any sequence without plateau, measured in a sitting position, without bronchodilator administration before the test, with an effort grade of "C" or lower and defined as a defective maneuver is discarded. The selected signals are rescaled by the corresponding BPTS factor, which is applied to each sequence to adjust the curve to lung conditions. Thus, final recordings are smoothed by an average moving filter of window size 10. A window size of 10 is translated into a 100 ms time span window. Fluctuations are removed, and the shapes and peaks of the spirometric curves are conserved. Equation 4.5 estimates the temporal window τ resulting from a period T on a certain window size N (Giordano and Knaflitz, 2019). The final result is two matching lists where every recording has a perfect match to the patient ID.

$$\tau = N \cdot T \tag{4.5}$$

Table 4.3: Conditions for the validation of spirometric sequences.

Condition	**Requirement**
Presence of plateau	Yes
Test body position	Standing
Pre-bronchodilator	Yes
Effort grade	"A" or "B"
Maneuver acceptability	Yes

The number of tests after filtering the sequences descends to 70234 recordings. Another condition that applies to spirometric examinations is that forced exhalation must last at least 6 seconds in adults to consider that the test is performed properly (Haynes, 2018). This minimum time is translated into a minimum curve length, which corresponds to 600 volume-time data points. Around 3.7 % of the filtered recordings are below this number of samples and 67639 sequences can be used.

4.3.3 Preservation of the Main Peak in the Spirometric Signals

In the clinical context, some parts of the curve have a greater influence on the outcome of the test. Peak Expiratory Flow (PEF) is reached after forced exhalation is completed and all air stored in the lungs is evacuated. This peak corresponds to the highest point on the spirometric curve, which is usually placed within the first seconds of the test. This specific point of the volume-time series reflects the TLC of the patient and his ability to perform a forced exhalation. The total shape of the PEF, from the beginning of the rise to the plateau, serves as an indicator of restriction or obstruction of the airways of the patients. The obstruction of the airways suggests the presence of respiratory conditions such as COPD or asthma (M R Miller et al., 2005a). Therefore, the trajectory generated in PEF could be extremely informative toward the patient's diagnosis.

A local maxima search algorithm is implemented to detect all peaks, restricting the search to peaks with a minimum amplitude of 2 L and a minimum distance within peaks of at least 100 samples, which is translated into a 1 second difference. The highest peak (PEF) is conserved. Another search begins around the peak to find the start and end of the curve segment containing the main peak curve. Since PEF has the tendency to be preceded by a sudden rise and a progressive decline, two windows of 80 samples and 120 samples are deployed to the left and right of the main peak. The objective then becomes to find the absolute minima within these intervals. These values will be assigned as the beginning and the end of the PEF curve. Figure 4.8 shows the identification of the peak in the first patient, whose original sequence is shown in Figure 4.7.

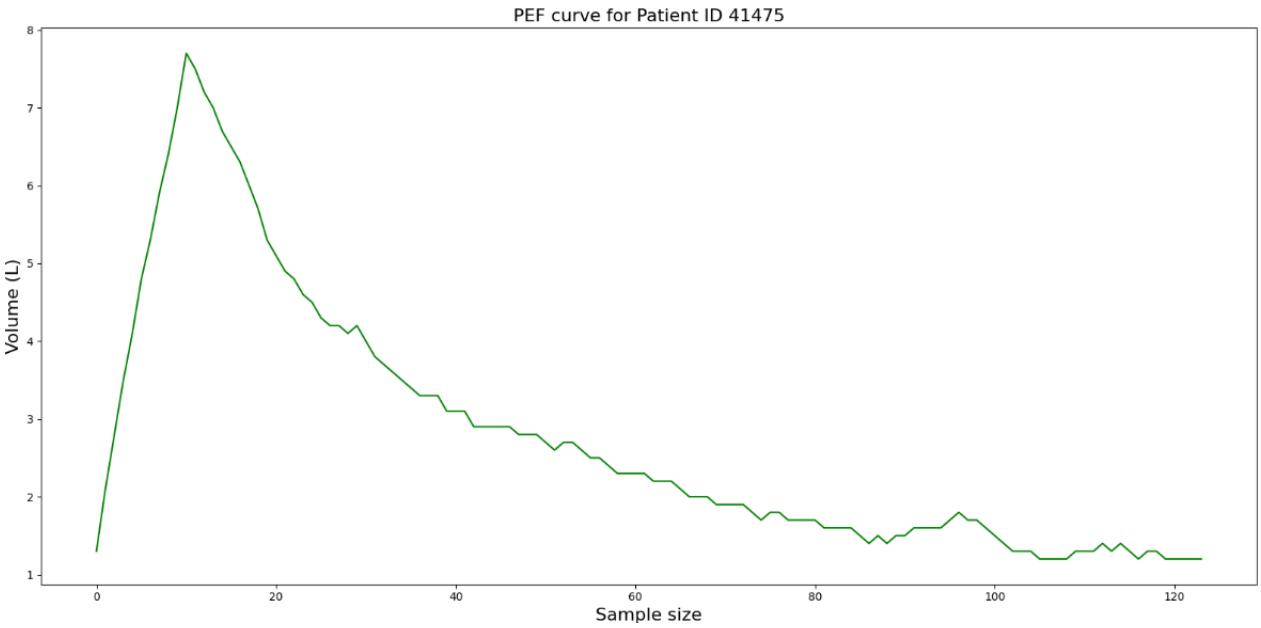

Figure 4.8: Main peak identification representation.

4.4 Methodological Framework for Approaches Development

4.4.1 Data Formatting and Splitting

Data splitting is performed by assigning 20 % of the data to the test set and using the rest to train the model. Parallel lists match the sequences to their corresponding patient ID's, for both the raw and the reduced peak recordings. Data are randomly assigned to training and test sets, setting a seed to ensure reproducible results later on. Sequences are chosen on the basis of their participant ID. To avoid sample leakage, sequences that come from the same individuals are not attributed to different sets. In doing so, the models will be tested in unseen patients.

As some models require equal sequence length; both peak and raw recordings are padded to the maximum sequence length. Short arrays are padded to the maximum length by adding zeros until the sequence ends. Thus, the beginning of the signal, which contains the PEF, is preserved. Although some studies highlight that pre-padding increases model performance (Dwarampudi and Subba Reddy, 2019), in the context of spirometry, the chances that the models learn from patterns in the given chronological order are greater when post-padding

is applied to the sequences.

Unsupervised learning techniques are characterized by using distance metrics to generate clusters from data. Normalizing the features ensures that all variables are equally influential in their contribution. Therefore, data are normalized via Min-Max standardization, which scales the features between values within [0,1]. This standardization method preserves the original data distribution, so that relevant patterns in the data are not disrupted (Wongoutong, 2024). Min-max standardization rescales every variable as shown in Equation 4.6, where the smallest value ($X_{\min}$) takes the value of 0, the largest value ($X_{\max}$)is equal to 1 and the rest is calculated proportionally between.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4.6}$$

4.4.2 Cluster-based Spirometric Data Exploration

The starting point is to explore how the data are naturally partitioned. The aim is to generate two well-differentiated groups between healthy and diseased patients. Several techniques,such as K-Means, DBSCAN, and Agglomerative clustering, are applied to both full and peak sequences. The implementation of these algorithms proceeds as follows.

- **K-Means** is configured to split the data between healthy and diseased patients. The initial centroids are randomly assigned and each data point is allocated to the cluster corresponding to the closest centroid. Centroids are iteratively updated as the mean value of all points in the cluster until the algorithm goes through all subset data points (Zhu et al., 2021). The Euclidian distance is employed as a metric to estimate the proximity of data points to clusters (Equation 4.7). Euclidian distance calculates the square root of the sum of the differences between every vector component x_i and y_i (Zhu et al., 2021).
- **Density-Based Spatial Clustering of Applications with Noise** (DBSCAN) is applied to spirometric recordings, expecting that groups of points packed together are assigned to the same cluster and distant points are defined as outliers. The hyperparameters of the model are the maximum distance between neighboring sequences and the minimum number of points to form a cluster. Core points share the minimum number of neighbors within a dense region, Border points possess less neighbors but are still found within the distance of the region, and Noise points correspond to outliers (Perafan-Lopez et al., 2022). DBSCAN aims to represent self-made clusters from the data to understand how feasible is the two-group model.

- **Agglomerative clustering** is initiated with all sequences encoded as single clusters. By means of the Euclidian distance, the proximity between clusters is computed. Then, the clusters are merged with each other following a given linkage criterion. For this project, Ward's method unifies clusters by minimizing variance within clusters (L. Yu et al., 2011). Clusters are merged iteratively until the desired number of groups (two in this case) is achieved.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.7}$$

The post-padded data sequences present 2310 and 200 dimensions for the full sequences and the peak data. This number of dimensions is too large to be graphically represented in the form of visible clusters. Thus, Principal Component Analysis (PCA) is applied to obtain the 2 main principal components that capture the highest amount of variability among data. In this scenario, PCA is used as a dimensionality reduction technique, which compresses the data, capturing the directions that maximize the variance (Groth et al., 2013).

4.4.3 Autoencoder and K-Means Clustering Approach

Deep-learning based autoencoders can be used for the detection of anomalies or characteristic patterns in one-dimensional time series data in medical applications (Nawaz and Ahmed, 2022). Retention of characteristic shapes and detection of anomalies in one-dimensional biomedical signals applied to unlabeled data in unsupervised learning approaches (Chadha et al., 2021).

Autoencoders are neural networks involved in unsupervised learning that allow to extract relevant representations from input data. The initial step is the encoding, which captures core patterns from data by representing the input vectors into a lower dimensional space. Unlike methods such as PCA, nonlinear activation functions enable modeling complex relationships among variables. The second phase is the decoding, where the algorithm attempts to reconstruct the compressed representation to the original input data. During the reconstruction, external noise is filtered and recognizable patterns are kept (Baig et al., 2023).In the present work, the purpose of the autoencoder is to generate a compressed representation of the spirometric sequences and introduce it inside the clustering algorithm. The same procedure is performed in both full and peak sequences. The sole distinction between datasets resides in the architecture of the autoencoder.

Before the model construction, data must be formatted in a compatible shape. All sequences are configured in a two-dimensional vector $(timesteps, features)$, where $features$ represents the volume records. The data enters the autoencoder passing through a succession of one-dimensional convolution layers, which apply a decreasing number of filters to the data. The activation function for every convolution layer is a Rectified Linear Unit (ReLU), which allows the model to learn from complex patterns introducing non-linearity to the network. ReLU activation function implements sparsity and improves computational efficiency in model training (Li et al., 2025). Equation 4.8 sets to 0 negative data entries, while preserving positive vector values.

$$\text{ReLU}(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0 \end{cases} \tag{4.8}$$

In between convolutions, one-dimensional pooling layers reduce the encoded sequence length by half, selecting segments of two consecutive time steps and keeping the greatest value. A final convolution layer is applied to generate the final encoded sequences, which are compressed representations of the original recordings for the entire and peak data.
Once the forward framework of the autoencoder is completed, compressed sequences enter the decoding phase to rebuild the original signal representation. A succession of convolution layers with increasing number of filters is interspersed with several up-sampling layers, producing the opposite effect than in the encoder structure. A cropping layer is added to match the exact input shape of the original sequences and a final one-dimensional convolution layer allows the final reconstruction. In this last convolution layer, ReLU activation is replaced by the sigmoid function (Equation 4.9), which outputs the final result in a range of [0,1] (O. A. M. López, A. M. López, and Crossa, 2022). In this context, the output values represent proportionally the similarity between the original and reconstructed sequences.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.9}$$

The model is compiled through the Adam optimizer, which is an extension of stochastic gradient descent (SGD). Adam optimizer is characterized by individually adapting the learning rate for each parameter and converges faster than conventional SGD (Kingma and Ba, 2015). Adam updates the model weights during training, minimizing the prediction error. In this case, the Mean Squared Error (MSE) is the selected loss function. MSE computes the square of the average difference between the reconstructed and original sequences (Equation 4.10) (Deng, Fu, and K. Yu, 2024).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.10}$$

The KerasTuner is the structure used to estimate the most efficient combination of hyperparameters for the autoencoder. During hyperparameter optimization, different configurations are explored on the number of filters in the convolution layers, the kernel size, and the learning rate in the optimizer. All possible values for hyperparameters are summarized in Table 4.4. The best combination of parameters is randomly searched within a predefined number of trials, in which each trial is trained only once. The minimum validation loss is adopted as the best model selection criterion.

Table 4.4: Hyperparameters combinations for 1D Autoencoder.

Hyperparameter	**Layer / Component**	**Possible Values**
filters1	Conv1D (Encoder 1)	32, 64, 128
kernel1	Conv1D (Encoder 1)	3, 5, 7
filters2	Conv1D (Encoder 2)	16, 32, 64
kernel2	Conv1D (Encoder 2)	3, 5, 7
filters3	Conv1D (Encoder 3)	8, 16, 32
kernel3	Conv1D (Encoder 3)	3, 5, 7
filters4	Conv1D (Encoded layer)	1, 4, 8
kernel4	Conv1D (Encoded layer)	3, 5, 7
filters5	Conv1D (Decoder 1)	8, 16, 32
kernel5	Conv1D (Decoder 1)	3, 5, 7
filters6	Conv1D (Decoder 2)	16, 32, 64
kernel6	Conv1D (Decoder 2)	3, 5, 7
filters7	Conv1D (Decoder 3)	32, 64, 128
kernel7	Conv1D (Decoder 3)	3, 5, 7
learning_rate	Adam Optimizer	0.01, 0.001, 0.0001

The one-dimensional autoencoder delivers a compressed representation of the original peak and the entire signal. This encoded layer is fed to a K means algorithm. In the current work, K-Means is set as the clustering model to create 2 groups that differentiate between disease and non-disease populations. The distance metric applied to K-Means is once again the Euclidean distance 4.7.

4.4.4 Reservoir Computing and K-Means Clustering

Reservoir computing models become a common approach for classification, prediction, or anomaly detection problems in biomedical time series. An input layer connects the input spirometry signals to the reservoir. The reservoir is composed of a fixed recurrent neural network with random connections between neurons. Due to the internal dynamics of the reservoir, time series are transformed into high informative characteristics towards the signals. The readout is the sole step of the algorithm that is trained, and it maps the dynamics of the reservoir to a desired output (Dale et al., 2019).Reservoir computing has been applied to other clinical fields for classification of biomedical time series, such as the classification of heart beats through electrocardiogram (ECG) signals (Hadaeghi, 2019) or electromyography (EMG) of hand gesture patterns for recognition of neuromorphic diseases (Garg et al., 2021).

In the present work, the reservoir is built based on a previous hyperparameter optimization that tunes the number of neurons, the leak rate, which is in charge of controlling the update of the neuron states and so balances the memory and reactivity of the neurons, and the spectral radius, which regulates the stability behavior of the reservoir dynamics. A large number of neurons presents richer dynamics within the reservoir, enabling a potential better performance in exchange for a higher computational cost and the risk of retaining some signal external noise. The leak rate controls the influence of previous states on how current input affects the neuron (Lukoševičius, 2012).

During training, all sequences within the training set are iteratively exposed to the reservoir. The dynamic states are conserved for every recording. Before every sample is introduced, the previous internal state processed by the reservoir is cleared. The resulting states of the reservoir are considered as the extracted features of the model. Evaluating the reservoir performance based on its dynamics allows to select the most efficient and parameters. Dynamics can be numerically translated by variance. The reservoir configuration providing the largest mean variance score (MVS) will be the model to take into account. Equation 4.11 reflects the MVS, where h_i refers to the reservoir states and N to the total number of samples. In addition, the variance is calculated as the mean of the squared differences between the states of neurons ($h_i(t)$) and the mean state ($\overline{h_i}$), as explained in Equation 4.12.

$$\text{Score} = \frac{1}{N}\sum_{i=1}^{N} \text{Var}(h_i) \tag{4.11}$$

$$\text{Var}(h_i) = \frac{1}{T}\sum_{t=1}^{T} \left(h_i(t) - \overline{h_i}\right)^2 \tag{4.12}$$

The same procedure as in Section 4.4.3 is then pursued. The neuron states are introduced to a K-Means algorithm which is configured to generate two clusters, differentiating diseased and healthy participants. Hence, test data are fed to the reservoir and the resulting states are classified based on formation of the previous clusters, to the closest centroid.

Chapter 5

Results

5.1 Experimental Setup

5.1.1 Clustering Models Application

The present study aims to apply the models to both raw and peak sequences and to compare their effectiveness in distinguishing between patients with obstructive respiratory disease and healthy participants. Both data sets are spread in a ratio of 4:1 in favor of the training portion. At the outset, K-Means, DBSCAN and Agglomerative clustering are applied to data to explore whether raw and peak sequences achieve to generate separate clusters for diseased and healthy individuals. Hence, the exact same data portion is fed to K-Means after being introduced to the autoencoder and the reservoir, to evaluate whether these models extract relevant signal characteristics and improve cluster generation. Another objective of the present work is to assess the diagnostic performance difference between employing the entire spirometry signal or a subset that contains the PEF. In the final phase, the study analyzes the model that produces better cluster formation across both data sets.

A subset of 20 individuals, half healthy and the other half have chronic obstructive disease, is used to identify whether the constructed clusters consistently differentiate between the two classes. The results of the models on this subset are representative towards the whole population of NHANES dataset. Therefore, the resulting clusters discriminate healthy cases from diseased patients. In order to obtain a clear graphical representation of clusters, PCA is applied to the model output. Thus, the two principal components that capture the highest variability among the data are plotted. Healthy participants are depicted with blue crosses, while diseased individuals are indicated with red markers. Figure 5.1 represents the volume-time curves for healthy and diseased samples.

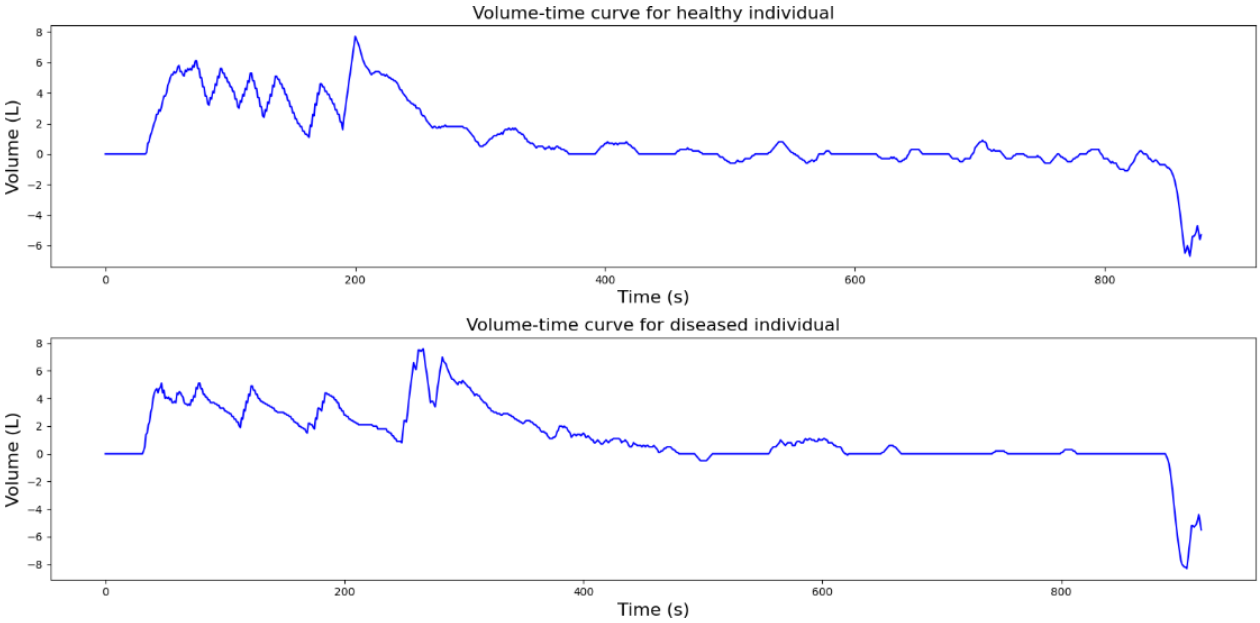

Figure 5.1: Spirometry curves for healthy and diseased patients.

Inspecting Figure 5.1 provides valuable information to understand the main differences between curves that identify healthy and diseased individuals. The most informative segment of the curve is the steepest peak, which corresponds to the PEF (Hoesterey et al., 2019). Regarding the ascent to that peak, the healthy individual reflects an abrupt increase in volume, while the affected patient presents a slower spike to the peak volume. This particular trend is related to the difficulty in the passage of air through the airways of the patient. The oscillations at the top of the sequence reveal that the patient with the disease struggles to exhale the air uniformly (Kakavas et al., 2021). After reaching maximum volume, the healthy participant curve exhibits a more progressive and smoother descent than the patient with obstructed airways. In summary, such a long and incomplete expiratory maneuver as the one performed by the second individual in Figure 5.1 is associated with EPOC and other obstructive diseases (Mochizuki et al., 2019).

5.1.2 Metrics for Performance Evaluation

The main tool used to evaluate the quality and consistency of the clusters is the Silhouette Score. This measurement indicates how well the model assigns points to their clusters and how the clusters differ from each other. To measure the quality of the clusters, the average individual silhouette scores of all points must be assessed. Negative values in the silhouette score highlight that the points are misclassified, values close to 0 determine that some points are very close to the other cluster and could be wrongly assigned, and values close to 1

ensure correct cluster assignment and formation (Zhao et al., 2018). Equation 5.1 describes the formula for computing the quality of the clusters through the Silhouette score, where $a(i)$ is the intra-cluster distance between a given point and the other points on the cluster and $b(i)$ refers to the average distance to the closest cluster, which states the mean distance between a given point and all points in the nearest neighbor cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i),\ b(i)\right)} \tag{5.1}$$

5.2 Clustering Analysis on Raw Spirometric Time Series

Within all clustering algorithms mentioned in Section 4.4.2, DBSCAN has not been successful in creating clusters from data, for both entire and peak sequences. Regarding K-Means and Agglomerative Clustering, two clusters arise from the models for both peak and full recordings. However, applying the models to the subset of 20 participants, the results do not show any correlation between belonging to a given cluster and different clinical outcomes. Healthy individuals and patients with the disease are mixed in the same classes. Testing on full sequences data, K-Means achieved a silhouette score of 0.12 while Agglomerative Clustering scored 0.097. A more accurate cluster separation is achieved on the PEF subset data, with a silhouette score of 0.319 for the K-Means algorithm and a score of 0.224 for Agglomerative Clustering. Figures 5.2 and 5.3 show K-Means results, respectively, applied to the test data in full and peak sequences.

Figure 5.2: K-Means clustering for complete time series data.

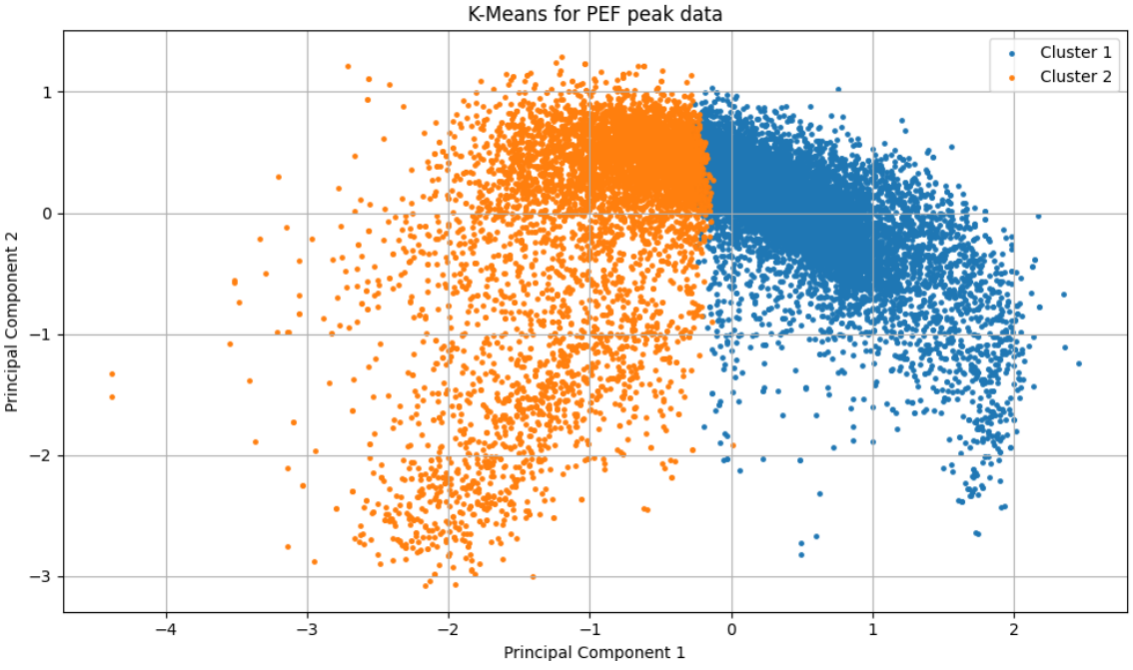

Figure 5.3: K-Means clustering for PEF peak data.

5.3 Model 1: Autoencoder and K-Means Results

The autoencoder is fitted to the training data, and the subset of hyperparameters from Table 4.4 that projects the best results is kept for the predictions of the sequences. The same procedure is followed for entire sequences and PEF peak data. Table 5.1 provides the optimal characterization of the hyperparameters for each data set.

Table 5.1: Optimal hyperparameters combination of the Autencoder for each data subset.

Hyperparameter	**Full sequences**	**Peak sequences**
filters1	128	64
kernel1	3	5
filters2	16	32
kernel2	3	3
filters3	8	16
kernel3	3	7
filters4	8	1
kernel4	3	3
filters5	32	8
kernel5	3	5
filters6	64	16
kernel6	5	5
filters7	32	32
kernel7	5	7
learning_rate	0.01	0.001

Thereafter, compressed representations from both models result in encoded vectors with shapes, respectively, of 289 and 25 dimensions for full and peak subsets. The compressed encoded data are fed to the K-Means algorithms and two clusters are obtained. The results differ between K-Means applied to the peak and the entire sequences. Cluster recognition is effectively achieved in the model applied to long spirometry time series and unhealthy patients are separated from healthy individuals, as shown in Figure 5.4. In contrast, the PEF sequences appear to be unable to distinguish between diseased and healthy patients, as can be seen in Figure 5.5. Regarding discrimination of generated clusters, the silhouette score for original recordings reports a value of approximately 0.16, while the score for compressed peak data is 0.531.

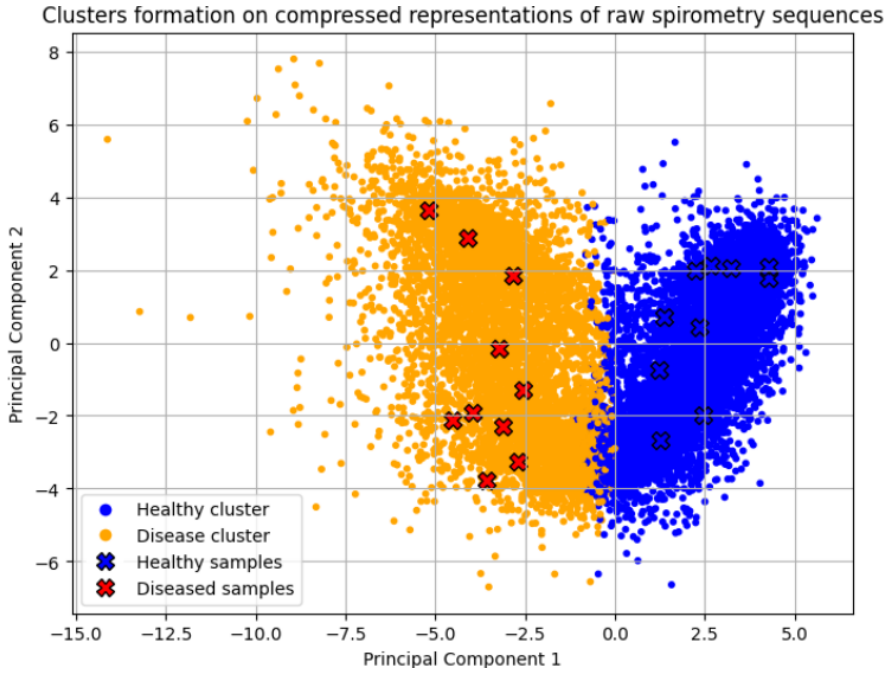

Figure 5.4: Clusters formation on compressed representations of raw spirometry sequences

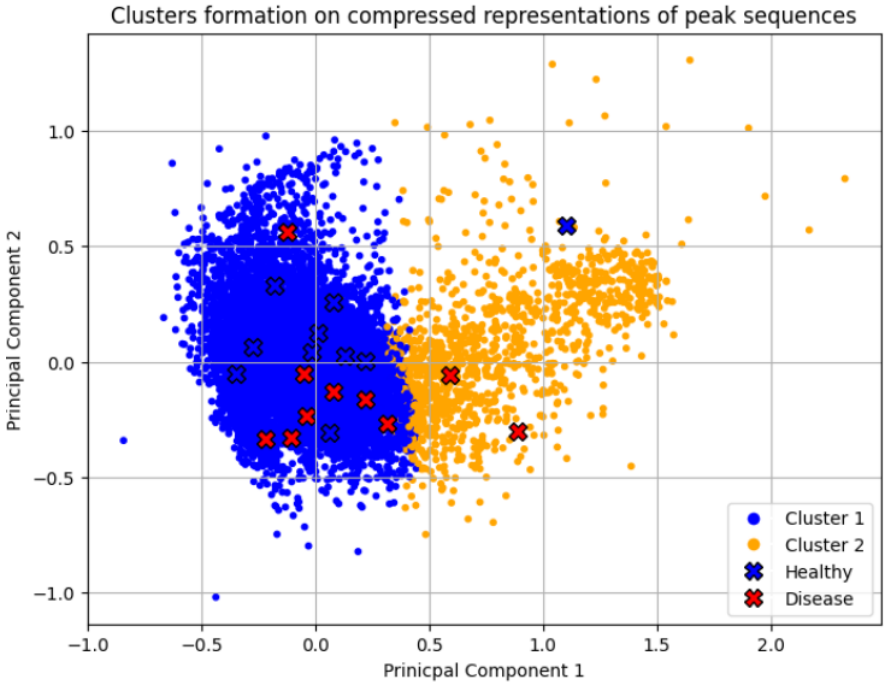

Figure 5.5: Clusters formation on compressed representations of peak sequences

5.4 Model 2: Reservoir and K-Means Results

The reservoir is fitted to the standardized training data and the combination of hyperparameters that provide the highest variability, so the richest set of extracted features, is selected. Table 5.2 shows the parameter setting that maximizes the training states obtained from the reservoir applied to the entire sequences and the peak PEF data.

Table 5.2: Optimal hyperparameters combination of the Reservoir for each data subset.

Reservoir model	**Number of neurons**	**Leaking Rate**	**Spectral Radius**
Full sequences	150	1.0	0.7
Peak sequences	200	1.0	1.0

The final states of the reservoirs represent a set of features extracted from the training data. These states present 150 features for entire spirometry recordings and 200 for the peak data. The states are then fed to K-Means to derive two clusters. Similarly to the autoencoder, Figure 5.6 demonstrates that the reservoir is able to differentiate between healthy and diseased cases for the raw sequences. However, it fails in generating reliable groups for peak data (Figure 5.7). The discriminative power between clusters appears to be significantly greater for peak data, with a silhouette score of 0.461, whereas in the entire spirometry sequences, the score drops to 0.073.

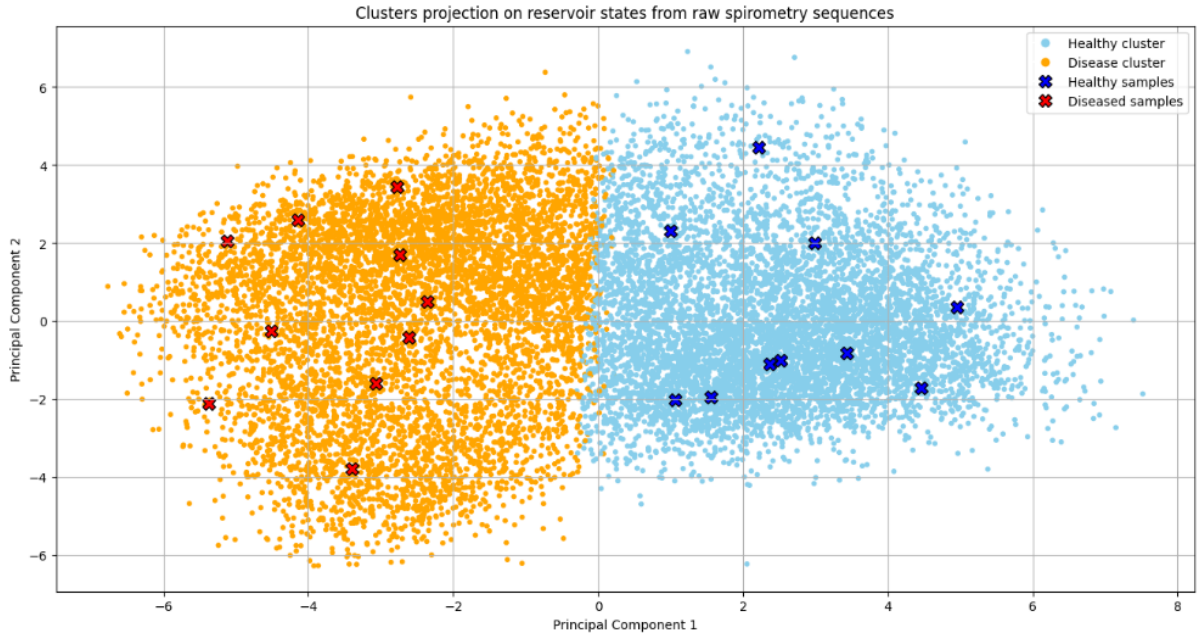

Figure 5.6: Clusters formation on reservoir states of raw spirometry sequences.

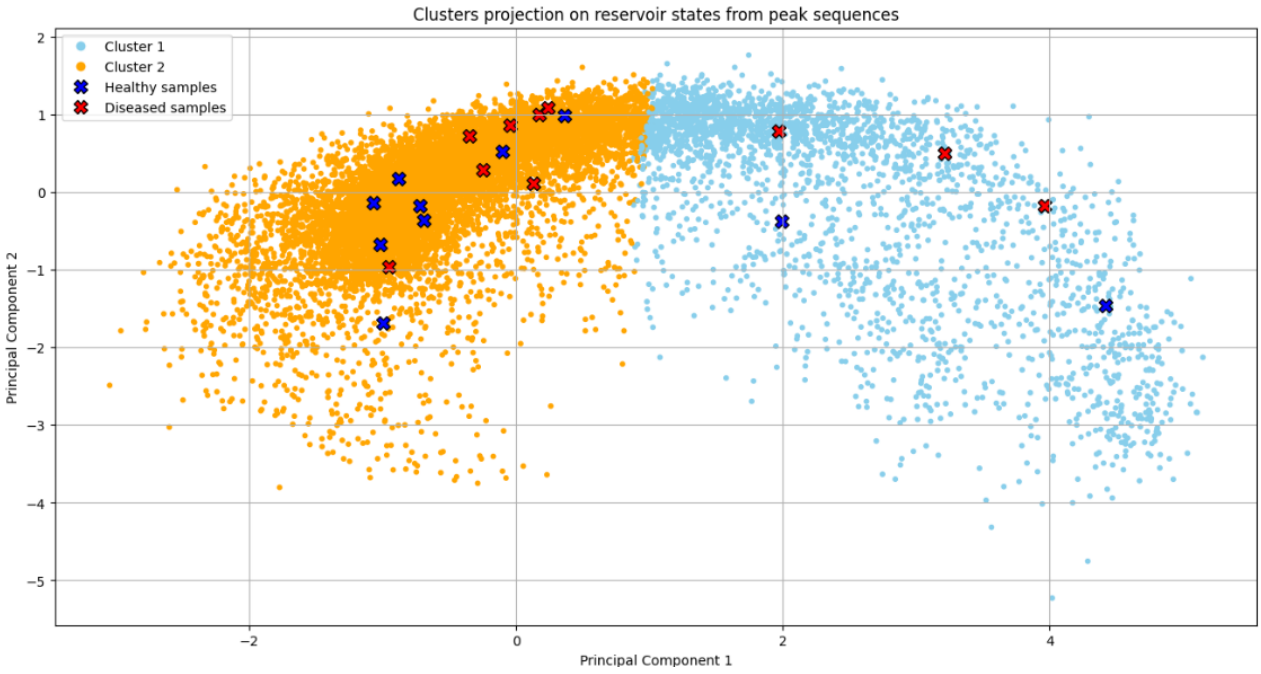

Figure 5.7: Clusters formation on reservoir states of peak sequences.

5.5 Analysis of Cluster Boundary Cases

Exploring undetermined cases that belong to the boundary generated between the two clusters provides valuable insight about what kind of behavior in spirometry sequences is shared in both healthy and diseased individuals. It is worth evaluating only the models that have clinical significance within the classes. These correspond to the algorithms applied to complete time series. Figure 5.8 provides an example of two individuals whose compressed representations of the autoencoder end at the frontier between the diagnosis groups. Both participants are located in the same area, suggesting that volume-time series should reflect similar patterns and shapes on the surface.

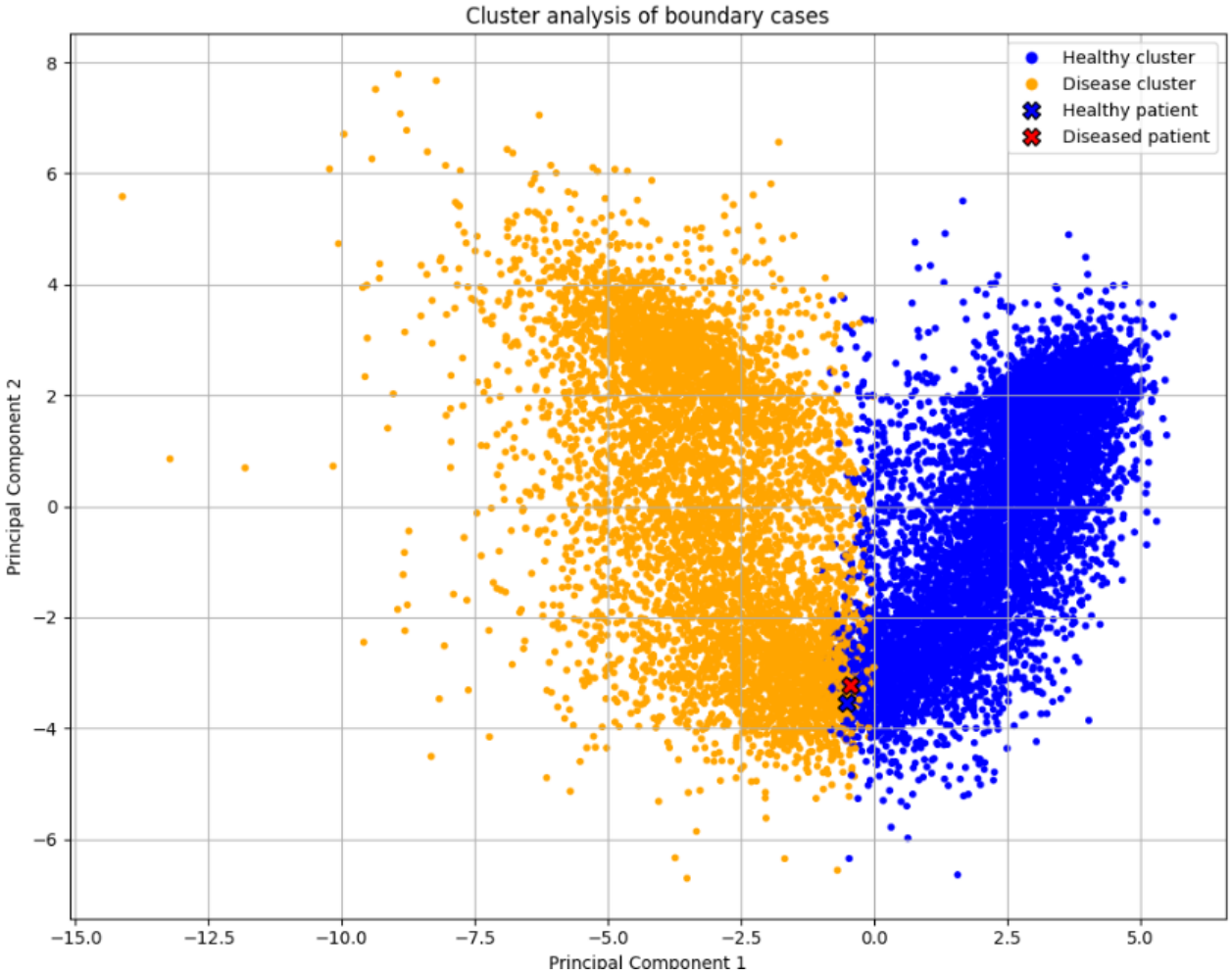

Figure 5.8: Representation of patients at cluster boundaries.

At first glance, it may seem that there are no significant differences in the spirometry curves corresponding to the boundary cases, shown in Figure 5.9. However, it is noticeable that the rise to the highest peak is moderately faster in the healthy record than for the patient, which can be diagnosed as sick. Furthermore, some signal fluctuations are perceived on the aforementioned rise in the curve of the diseased patient. These oscillations are one of the main indicators of possible air obstruction problems. In addition to that, the descent is visibly slower in the second graph than in the healthy individual representation.

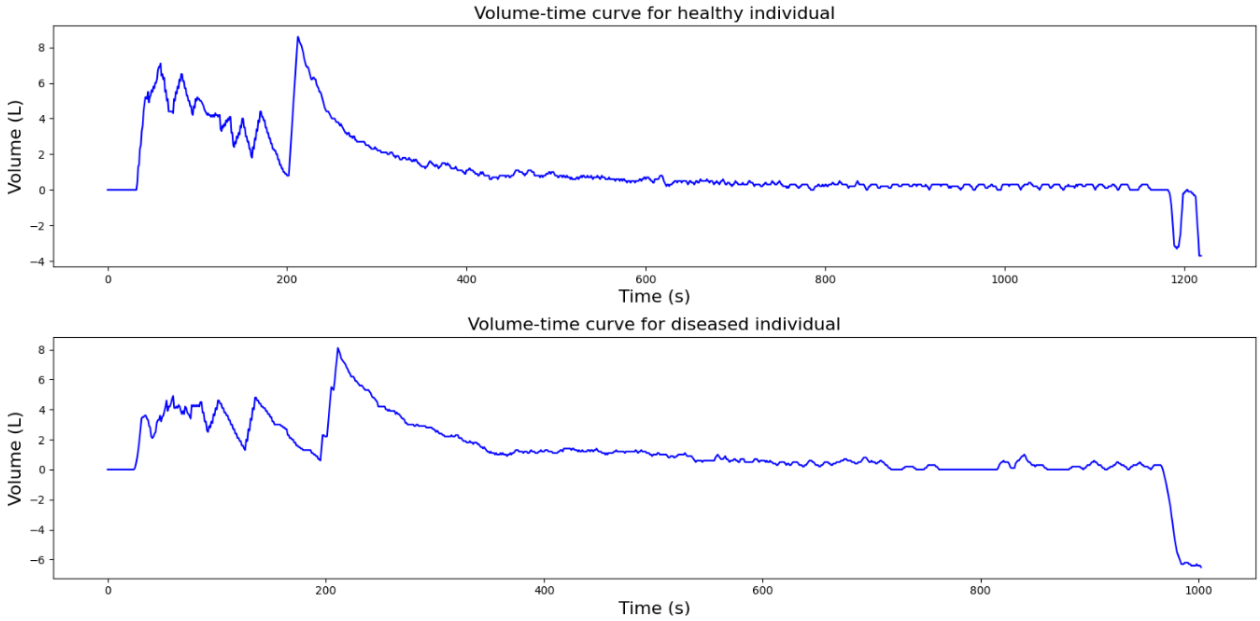

Figure 5.9: Spirometry curves for patients on clusters boundaries.

Chapter 6

Discussion

6.1 Clustering Algorithm Selection Analysis

Before model estimation, the current study demonstrates that K-Means is the clustering algorithm that provides a better cluster generation for spirogram volume-time series, compared to other methods. DBSCAN typically struggles on datasets that contain different data point densities and is not a recommended option for high-dimensional data (Yongyu Wang, 2024). These characteristics are perfectly recognizable in the spirometry data. DBSCAN does not conform to time series structured as high-dimensional data. Similar concerns are encountered with Agglomerative clustering, which unlike DBSCAN, is sensitive to noise and atypical values. Moreover, hierarchical clustering represents a high computational expense due to the way it proceeds to create the clusters (Balcan, Liang, and Gupta, 2014). K-means arises as the most suitable technique, as shown in Figures 5.2 and 5.3, where two clusters of spherical shape are generated from the complete and peak data subsets. The advantages of this algorithm are that they perform fast and efficiently in large datasets and are convenient for equal-sized and compact classes (Wani, 2024). Furthermore, the number of desired clusters is a hyperparameter of the model that can be fixed prior to cluster formation, which enables restricting the analysis to disease and healthy classes. Table 6.1 displays the silhouette score differences between hierarchical and K-means clustering, demonstrating that the second algorithm provides a high-level spread of groups.

Table 6.1: Silhouette scores for clustering algorithms applied to full and peak sequences.

Clustering Algorithm	**Full Sequence**	**PEF Subset**
K-Means	0.12	0.319
Agglomerative Clustering	0.097	0.224

6.2 Models Comparison

Similar results are obtained by applying the K-means algorithm to the compressed representations of the autoencoder (Figure 5.4) and to the states extracted from the ESN (Figure 5.6). Both models show some advantages and disadvantages while working with sequential temporal data, which can be listed as follows.

- **Autoencoder**
  - The latent space (compressed input data) is able to capture characteristic trends, high-level patterns, anomalies, and unusual details among the data (Cai et al., 2023).
  - Ideally, it is designed to provide compressed sequences that can be fed to consecutive models, such as clustering algorithms (Sepehr Maleki, Sasan Maleki, and Jennings, 2021).
  - It allows a wide variety of architectures depending on the final purpose of the analysis (Sepehr Maleki, Sasan Maleki, and Jennings, 2021).
  - It requires complex and time-consuming training, where gradients risk to vanish in long sequences and presents the possibility of overfitting data (Yang et al., 2025).
- **Reservoir**
  - The extracted states manage to capture the temporal dependencies that encode the dynamics within the data (Sakemi et al., 2024).
  - It avoids backpropagation because it only needs to process the output weights, making training time efficient (Lukoševičius, 2012).
  - It presents a lower risk of overfit (Lukoševičius, 2012).
  - It is really sensitive to hyperparameters, experiencing difficulties in capturing long-term pattern associations within sequences, which limits its ability to retain complex dependencies (Verstraeten et al., 2007).

Figures 5.4 and 5.6 confirm that both models allocate healthy and diseased cases within their respective clusters. The reservoir provides more dispersion in the spectrum of points than the autoencoder encoded representations. The fact that subjects are placed similarly within the clusters indicates that temporal dependencies have been effectively retained in both approaches. Considering that volume-time curves present the vast amount of differences in the PEF segment (for example, in patients depicted in Figure 5.1), models achieve a low-dimensional representation that has an impact on the clinical decision that assigns subjects as healthy or diseased.

The initial states of the reservoir that are prone to control the primary dynamics of time series are called transients. Transients are randomly initialized, which means that they are not valid representations of the input sequences (Yildiz, Jaeger, and Kiebel, 2012). The common way to proceed is to remove transients to avoid introducing biased temporal dependencies at the beginning of the sequence. It is reasonable to argue that the reservoir models would perform better removing the initial states that are fed to K-Means. However, similar results are obtained excluding the transients in complete sequences analysis. This is probably due to the fact that the beginning of signals does not contain relevant discriminatory patterns. In case of PEF sequences, eliminating transients from peak data conditions the retention of valuable patterns. After deleting transients, the performance of the reservoir on peak data remains mediocre.

The solutions differ mainly in the complexity, as the Autoencoders require the design of a high-level architecture which combines convolutions with down-sampling and up-sampling layers, including hyperparameter tuning refinement, while the ESN only relies on several parameters and provides a sparse recurrent neural network as the core structure. In addition, ESN are less time-consuming than one-dimensional convolution layers, as the only layer that needs to be trained is the output section of the model. In addition to that, K-means reveal that compressed representations of the Autoencoder provide an augmented inter-class variance compared to the reservoir states, as described in Table 6.2.

Table 6.2: Models silhouette scores.

Model	**Full Sequence**	**PEF subset**
Reservoir states	0.073	0.461
Autoencoder compressed data	0.160	0.531

Regarding the models applied to the PEF segment of the recordings, both alternatives fail to generate clinically meaningful groups, as depicted in Figures 5.5 and 5.7. Several interpretations can be hypothesized to explain these results. Intermediate subsequences that precede or succeed PEF that contain additional discriminatory power toward the final clinical outcome are lost by uniquely conserving peak curves. In case of the autoencoder, this model is based on complete temporal evolution of the signals to capture the totality of dynamics and dependencies within the data (Mienye and T. G. Swart, 2025). On the other hand, the peak extractor may be deficient between samples.

Another source of error may be the reduction in variability that entails a smaller range of available values. In this way, models reduce their ability to capture differentiating patterns in data (Angelike and Musch, 2024). Last but not least, although the noise reduction is applied to signals, some artifacts are always kept in the peaks. Peaks are areas in the signal

that normally concentrate noise. In shorter sequences, models are more sensitive to these interferences, which differ between volume-time series (Rohr et al., 2024).

Regarding the separability of clusters, applying K-means to PEF segments provides better results (Table 6.2) than in full sequences. As expected, lower-dimensional fragments of the data benefit distance metrics in clustering algorithms. The length of the segments that constitute PEF data is shorter than that of the raw volume spirograms. This reduction in dimensionality is translated into smaller distance vectors, which are prone to magnifying cluster differences.

Some patients are found within the boundaries of the clusters. In these cases, inspecting the volume-time spirometry sequences can make the difference. For example, the scenario presented in Figure 5.8 can be resolved by in-depth inspection of the subject spirograms, shown in Figure 5.9. Therefore, the solutions appear to be effective for patients affected by severe respiratory disease. However, volume recordings showing slight fluctuations related to restrictive and obstructive diseases remain on the frontier between healthy and diseased individuals. Thus, a detailed exploration of the spirometry sequences, as described in Section 5.5, is needed to understand crucial differences in patients.

Chapter 7

Conclusions

Deep learning based hybrid approaches such as autoencoders and reservoir computing combined with clustering algorithms effectively classify subjects under suspicion of restrictive or obstructive respiratory conditions. Traditional diagnostic methodologies, which require visual inspection and calculation of additional lung function parameters, can be complemented with solutions involving unsupervised learning techniques. Thus, it has been proven that the first stage of the detection of severe respiratory disease can be assessed by means of the proposed solutions.

The volume time series of the spirometry trials must be recovered under the same experimental conditions. The administered medication and the anatomical position of the patient, combined with operational requirements such as signal plateau achievement, correct maneuver, and effort of the subject, are common factors that must be met in spirometry tests to provide quality data and obtain a reliable clinical result. The length of the sequences is directly correlated with the quality of the spirometry examination. Thus, it is important to filter the sequences by all these factors before fitting the models to the data. Pre-processing and denoising spirograms before feeding them to models prevents the algorithms from learning undesired patterns which would remain on data.

Hybrid approaches that combine compressed data representations from autoencoders and reservoir states with K-means clustering succeed in generating clusters that distinguish between patients with chronic respiratory disorders and healthy individuals. K-means outperforms other clustering alternatives as it adapts to spirometry recordings type-data, providing time-efficient and controllable cluster formation. Latent representations of the autoencoders encode the input volume-time series and capture relevant temporal patterns. In the same way, Echo State Networks capture temporal dependencies in a more robust and time-efficient way, but with a lower inter-class variances than the first model. Therefore, both models seem to be valid in terms of performance.

With respect to the different data subsets, total recordings appear to properly learn the temporal dependencies among sequence reconstructions or extracted features and separate healthy and diseased cases with improvable cluster separability but consistent clinical significance. Otherwise, peak data subset lose discriminatory patterns in exchange for increasing inter-variability of clusters. However, this approach is futile as no clinical significance is achieved between the groups.

The solutions provided are promising for patients with spirograms that reflect severe obstructions or restrictions of the respiratory tract. However, the cases in which there may be a regression or progression toward the disease remain unclear at the groups boundary. These are the scenarios in which the combination of these deep-learning approaches and human clinical intervention and interpretation is necessary.

Looking beyond, several improvements could be implemented to the current project pipeline and to the hybrid models:

- Regarding alternatives applied to peak data, develop an accurate peak detector and establish a automatization in the management of transients could be the key measure to improve the models performance.
- Apply the models to labeled spirometric sequences to confirm whether the final approaches succeed in detecting restrictive or obstructive patterns and differentiate between healthy and diseased subjects.
- Deploy a similar approach being able to, not only distinguish between disease status, but also provide an accurate classification between restrictive and obstructive respiratory conditions.

References

American Thoracic Society (1999). "Dyspnea. Mechanisms, assessment, and management: a consensus statement". In: *American Journal of Respiratory and Critical Care Medicine* 159.1, pp. 321–340. DOI: 10.1164/ajrccm.159.1.ats898. URL: https://pubmed.ncbi.nlm.nih.gov/9872857.

Angelike, Tim and Jochen Musch (2024). "A comparative evaluation of measures to assess randomness in human-generated sequences". In: *Behavior Research Methods* 56. Original Manuscript, Accepted: 5 June 2024 / Published online: 1 July 2024, pp. 7831–7848. DOI: 10.3758/s13428-024-02456-7. URL: https://doi.org/10.3758/s13428-024-02456-7.

Al-Ashkar, Feyrouz, Reena Mehra, and Peter J. Mazzone (2003). "Interpreting pulmonary function tests: Recognize the pattern, and the diagnosis will follow". In: *Cleveland Clinic Journal of Medicine* 70.10, pp. 866–881. DOI: 10.3949/ccjm.70.10.866.

Augustin, Ingrid M. L. et al. (Sept. 2018). "The respiratory physiome: Clustering based on a comprehensive lung function assessment in patients with COPD". In: *PLOS ONE* 13.9, e0201593. DOI: 10.1371/journal.pone.0201593.

Azman, Josip et al. (2006). "Correlation and regression". In: *Acta Medica Croatica* 60.Suppl 1. [Article in Croatian], pp. 81–91.

Backman, Helena et al. (2016). "Restrictive spirometric pattern in the general adult population: Methods of defining the condition and consequences on prevalence". In: *Respiratory Medicine* 120. Epub 2016 Oct 12, pp. 116–123. DOI: 10.1016/j.rmed.2016.10.005. URL: https://pubmed.ncbi.nlm.nih.gov/27817808/.

Baig, Yasa et al. (Dec. 2023). "Autoencoder neural networks enable low dimensional structure analyses of microbial growth dynamics". In: *Nature Communications* 14.1, p. 7937. DOI: 10.1038/s41467-023-43455-0.

Balcan, Maria-Florina, Yingyu Liang, and Pramod Gupta (2014). "Robust Hierarchical Clustering". In: *arXiv preprint arXiv:1401.0247*. Version 2, last revised 13 Jul 2014. DOI: 10.48550/arXiv.1401.0247. URL: https://doi.org/10.48550/arXiv.1401.0247.

Barnes, T. and L. Fromer (2011). "Spirometry use: Detection of chronic obstructive pulmonary disease in the primary care setting". In: *Clinical Interventions in Aging* 6, pp. 47–52.

Barreiro, Timothy J. and Irene Perillo (2004). "An Approach to Interpreting Spirometry". In: *American Family Physician* 69.5, pp. 1107–1115. URL: https://www.aafp.org/pubs/afp/issues/2004/0301/p1107.html.

Bertels, Xander et al. (Feb. 2024). "Clinical relevance of lung function trajectory clusters in middle-aged and older adults". In: *ERJ Open Research* 10.1, pp. 00793–2023. DOI: 10.1183/23120541.00793-2023.

Bhatt, Surya P. et al. (2023). "FEV1/FVC severity stages for chronic obstructive pulmonary disease". In: *American Journal of Respiratory and Critical Care Medicine* 208.6, pp. 676–684. DOI: 10.1164/rccm.202303-0450OC.

Bickel, Scott et al. (Sept. 2014). "Impulse oscillometry: Interpretation and practical applications". In: *Chest* 146.3, pp. 841–847. DOI: 10.1378/chest.13-1875.

Cai, Borui et al. (2023). "Hybrid variational autoencoder for time series forecasting". In: *Knowledge-Based Systems* 281. Published 3 December 2023, p. 111079. DOI: 10.1016/j.knosys.2023.111079. URL: https://doi.org/10.1016/j.knosys.2023.111079.

Centers for Disease Control and Prevention (Aug. 2024). *What NHANES Data Have Achieved*. https://www.cdc.gov/nchs/nhanes/publications/data-accomplishments.html. Accessed on 14 Aug 2025.

Chadha, Gavneet Singh et al. (Aug. 2021). "Deep Convolutional Clustering-Based Time Series Anomaly Detection". In: *Sensors (Basel)* 21.16, p. 5488. DOI: 10.3390/s21165488.

Chen, Jing et al. (2020). "Prediction models for pulmonary function during acute exacerbation of chronic obstructive pulmonary disease". In: *Physiological Measurement* 41.12, p. 125010. DOI: 10.1088/1361-6579/abc792.

Cohen, Robert A. C., Aiyub Patel, and Francis H. Y. Green (2008). "Lung disease caused by exposure to coal mine and silica dust". In: *Seminars in Respiratory and Critical Care Medicine* 29.6. Epub 2009 Feb 16, pp. 651–661. DOI: 10.1055/s-0028-1101275. URL: https://pubmed.ncbi.nlm.nih.gov/19221963/.

Cukic, Vesna et al. (2013). "Asthma and Chronic Obstructive Pulmonary Disease (COPD) – Differences and Similarities". In: *Medical Archives* 67.5, pp. 291–294. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3633485/.

Dale, Matthew et al. (2019). "A substrate-independent framework to characterize reservoir computers". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2228. Received: 23 October 2018, Accepted: 15 May 2019, p. 20180723. DOI: 10.1098/rspa.2018.0723. URL: http://dx.doi.org/10.1098/rspa.2018.0723.

David, Sharoon, Jennifer Goldin, and Christopher W. Edwards (2024). *Forced Expiratory Volume*. Last Update: October 14, 2024. StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/books/NBK560526/.

Deng, Lu, Sheng Fu, and Kai Yu (Feb. 2024). "Bias and mean squared error in Mendelian randomization with invalid instrumental variables". In: *Genetic Epidemiology* 48.1. Epub 2023 Nov 16, pp. 27–41. DOI: 10.1002/gepi.22541.

Di Dio, Riccardo et al. (2021). "Spirometry-Based Airways Disease Simulation and Recognition Using Machine Learning Approaches". In: *Learning and Intelligent Optimization (LION 2021)*, pp. 98–112. DOI: 10.1007/978-3-030-92121-7_8.

Dwarampudi, Mahidhar and N V Subba Reddy (Mar. 2019). "Effects of padding on LSTMs and CNNs". In: *arXiv preprint arXiv:1903.07288*. DOI: 10.48550/arXiv.1903.07288. URL: https://arxiv.org/abs/1903.07288.

Fallahtafti, Farahnaz, Shane R Wurdeman, and Jennifer M Yentes (Aug. 2021). "Sampling rate influences the regularity analysis of temporal domain measures of walking more than spatial domain measures". In: *Gait & Posture* 88, pp. 122–127. DOI: 10.1016/j.gaitpost.2021.06.014.

Gadgil, Soham, Joshua Galanter, and Mohammadreza Negahdar (Nov. 2024). "Transformer-based time-series biomarker discovery for COPD diagnosis". In: *arXiv preprint arXiv:2411.08731*. Submitted on 13 Nov 2024.

Garg, Nikhil et al. (July 2021). "Signals to Spikes for Neuromorphic Regulated Reservoir Computing and EMG Hand Gesture Recognition". In: *Proceedings of the International Conference on Neuromorphic Systems 2021 (ICONS '21)*, p. 8. DOI: 10.1145/3477145.3477267. URL: https://doi.org/10.1145/3477145.3477267.

Giordano, Noemi and Marco Knaflitz (Apr. 2019). "A Novel Method for Measuring the Timing of Heart Sound Components through Digital Phonocardiography". In: *Sensors (Basel)* 19.8, p. 1868. DOI: 10.3390/s19081868.

Giri, Paresh C. et al. (2021). "Application of Machine Learning in Pulmonary Function Assessment: Where Are We Now and Where Are We Going?" In: *Frontiers in Physiology* 12, p. 678541. DOI: 10.3389/fphys.2021.678541. URL: https://doi.org/10.3389/fphys.2021.678541.

Graham, Brian L. et al. (2019). "Standardization of Spirometry 2019 Update. An Official American Thoracic Society and European Respiratory Society Technical Statement". In: *Am J Respir Crit Care Med* 200.8, e70–e88. DOI: 10.1164/rccm.201908-1590ST.

Groth, Detlef et al. (2013). "Principal components analysis". In: *Methods in Molecular Biology*. Vol. 930. Humana Press, pp. 527–547. DOI: 10.1007/978-1-62703-059-5_22.

Hadaeghi, Fatemeh (2019). "Computing Models for Patient-Adaptable ECG Monitoring in Wearable Devices". In: *Journal of Biomedical Signal Processing and Control*. Preprint submitted July 24, 2019.

Haynes, Jeffrey M. (2018). "Basic spirometry testing and interpretation for the primary care provider". In: *Can J Respir Ther* 54.4, 10.29390/cjrt-2018–017. DOI: 10.29390/cjrt-2018-017.

Hoesterey, Daniel et al. (2019). "Spirometric Indices of Early Airflow Impairment in Individuals at Risk of Developing COPD: Spirometry Beyond FEV1/FVC". In: *Respiratory Medicine* 156. Author manuscript; available in PMC: 2020 Sep 1, pp. 58–68. DOI: 10.1016/j.rmed.2019.08.004.

Ioachimescu, Octavian C., James K. Stoller, and Francisco Garcia-Rio (2020). "Area under the expiratory flow-volume curve: Predicted values by artificial neural networks". In: *Scientific Reports* 10, p. 16624. DOI: 10.1038/s41598-020-73925-0.

Kakavas, Sotirios et al. (2021). "Pulmonary function testing in COPD: looking beyond the curtain of FEV1". In: *npj Primary Care Respiratory Medicine* 31.23. DOI: 10.1038/s41533-021-00236-w. URL: https://doi.org/10.1038/s41533-021-00236-w.

Karuppiah, Rohini and G. Suseendran (Nov. 2016). "Aggregated K means clustering and decision tree algorithm for spirometry data". In: *Indian Journal of Science and Technology* 9.44. DOI: 10.17485/ijst/2016/v9i44/103107.

Kim, Hae-Young (Mar. 2017). "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test". In: *Restorative Dentistry and Endodontics* 42.2, pp. 152–155. DOI: 10.5395/rde.2017.42.2.152.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. arXiv:1412.6980 [cs.LG]. San Diego, CA. URL: https://doi.org/10.48550/arXiv.1412.6980.

Koegelenberg, C. F. N., F. Swart, and E. M. Irusen (2012). "Guideline for office spirometry in adults, 2012". In: *S Afr Med J* 103.1, pp. 52–62. DOI: 10.7196/samj.6197.

Kristensen, Kris et al. (Aug. 2023). "Using random forest machine learning on data from a large, representative cohort of the general population improves clinical spirometry references". In: *The Clinical Respiratory Journal* 17.8, pp. 819–828. DOI: 10.1111/crj.13662.

Lalloo, U G, M R Becklake, and C M Goldsmith (1991). "Effect of Standing Versus Sitting Position on Spirometric Indices in Healthy Subjects". In: *Respiration*. DOI: 10.1159/000195911.

Li, Jiayun et al. (Oct. 2025). "From ReLU to GeMU: Activation functions in the lens of cone projection". In: *Neural Networks* 190. Epub 2025 Jun 3, p. 107654. DOI: 10.1016/j.neunet.2025.107654.

López, Osval Antonio Montesinos, Abelardo Montesinos López, and Jose Crossa (2022). "Multivariate Statistical Machine Learning Methods for Genomic Prediction". In: *Fundamentals of Artificial Neural Networks and Deep Learning*. Springer, pp. 379–425. DOI: 10.1007/978-3-030-89010-0_15. URL: https://doi.org/10.1007/978-3-030-89010-0_15.

Lukoševičius, Mantas (2012). "A Practical Guide to Applying Echo State Networks". In: *Neural Networks: Tricks of the Trade, Reloaded*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer, pp. 659–686. ISBN: 978-3-642-35288-1. DOI: 10.1007/978-3-642-35289-8_36. URL: https://doi.org/10.1007/978-3-642-35289-8_36.

Mac, Amanda, Tong Xu, Joyce K Y Wu, et al. (2022a). "Deep learning using multilayer perception improves the diagnostic acumen of spirometry: a single-centre Canadian

study". In: *BMJ Open Respiratory Research* 9.1, e001396. DOI: 10.1136/bmjresp-2022-001396.

Mac, Amanda, Tong Xu, Joyce K. Y. Wu, et al. (2022b). "Deep learning using multilayer perception improves the diagnostic acumen of spirometry: a single-centre Canadian study". In: *BMJ Open Respiratory Research* 9.1, e001396. DOI: 10.1136/bmjresp-2022-001396. URL: https://doi.org/10.1136/bmjresp-2022-001396.

Maleki, Sepehr, Sasan Maleki, and Nicholas R. Jennings (2021). "Unsupervised anomaly detection with LSTM autoencoders using statistical data-filtering". In: *Applied Soft Computing* 108. Published September 2021, p. 107443. DOI: 10.1016/j.asoc.2021.107443. URL: https://doi.org/10.1016/j.asoc.2021.107443.

Mannino, David M, Enrique Diaz-Guzman, and Sonia Buist (Oct. 2011). "Pre- and Post-bronchodilator Lung Function as Predictors of Mortality in the Lung Health Study". In: *Respiratory Research* 12.1, p. 136. DOI: 10.1186/1465-9921-12-136.

Maree, D. M. et al. (2022). "Position statement for adult and paediatric spirometry in South Africa: 2022 update". In: *Afr J Thorac Crit Care Med* 28.4, 10.7196/AJTCCM.2022.v28i4.287. DOI: 10.7196/AJTCCM.2022.v28i4.287.

Martinez-Pitre, Pedro J., Bhanusivakumar R. Sabbula, and Marco Cascella (2025). *Restrictive Lung Disease*. StatPearls [Internet], updated 2023 Jul 25. Treasure Island, FL: StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/books/NBK560880/.

Masekela, R. et al. (2013). "Paediatric spirometry guideline of the South African Thoracic Society: Part 1". In: *South African Medical Journal* 103.12 Suppl 2, pp. 1036–1041. DOI: 10.7196/samj.7239.

Matabuena, Marcos, Aritra Ghosal, et al. (Feb. 2023). "Predicting Distributions of Physical Activity Profiles in the NHANES Database Using a Partially Linear Fréchet Single Index Model". In: *arXiv preprint arXiv:2302.07815*. Submitted on 15 Feb 2023; last revised 9 Mar 2025 (v2).

Matabuena, Marcos and Alexander Petersen (Apr. 2021). "Distributional Data Analysis of Accelerometer Data from the NHANES Database Using Nonparametric Survey Regression Models". In: *arXiv preprint arXiv:2104.00986*. Submitted on 2 Apr 2021; last revised 20 Jan 2022 (v2).

Mei, Shuhao et al. (2025). "Deep learning for detecting and early predicting chronic obstructive pulmonary disease from spirogram time series". In: *npj Systems Biology and Applications* 11.1, p. 18. DOI: 10.1038/s41540-025-00489-y.

Mienye, Ibomoiye Domor and Theo G. Swart (2025). "Deep Autoencoder Neural Networks: A Comprehensive Review and New Perspectives". In: *Archives of Computational Methods in Engineering*. Received: 26 June 2024 / Accepted: 28 February 2025. DOI: 10.1007/s11831-025-10260-5. URL: https://doi.org/10.1007/s11831-025-10260-5.

Miller, M R et al. (Aug. 2005a). "Standardisation of Spirometry". In: *European Respiratory Journal* 26.2, pp. 319–338. DOI: 10.1183/09031936.05.00034805.

Miller, M. R. et al. (2005b). "General considerations for lung function testing". In: *European Respiratory Journal* 26.1. ATS/ERS Task Force Review, pp. 153–161. DOI: 10.1183/09031936.05.00034505.

Miyoshi, Seigo et al. (July 2020). "Prediction of Spirometric Indices Using Forced Oscillometric Indices in Patients with Asthma, COPD, and Interstitial Lung Disease". In: *International Journal of Chronic Obstructive Pulmonary Disease* 15, pp. 1565–1575. DOI: 10.2147/COPD.S250080.

Mochizuki, Fumi et al. (2019). "The Concavity of the Maximal Expiratory Flow–Volume Curve Reflects the Extent of Emphysema in Obstructive Lung Diseases". In: *Scientific Reports* 9, p. 13159. DOI: 10.1038/s41598-019-49591-2. URL: https://doi.org/10.1038/s41598-019-49591-2.

Moon, Jinyoung and Yongseok Mun (Jan. 2025). "Construction of the Cancer Patients' Database Based on the US National Health and Nutrition Examination Survey (NHANES) Datasets for Cancer Epidemiology Research". In: *BMC Medical Research Methodology* 25.1, p. 17. DOI: 10.1186/s12874-025-02478-5.

Moreno Mendez, Rosaly et al. (2024). "Artificial Intelligence Applied to Forced Spirometry in Primary Care". English, Spanish. In: *Open Respiratory Archives* 6.Suppl 2, p. 100313. DOI: 10.1016/j.opresp.2024.100313.

National Guideline Alliance (UK) (Oct. 2017). *NICE Guideline No. 78*. London.

Nawaz, Menaa and Jameel Ahmed (2022). "Cloud-based healthcare framework for real-time anomaly detection and classification of 1-D ECG signals". In: *PLoS One* 17.12, e0279305. DOI: 10.1371/journal.pone.0279305.

Ong-Salvador, Rachel, Pierantonio Laveneziana, and Franciscus de Jongh (2024). "ERS/ATS Global Lung Function Initiative normal values and classifying severity based on z-scores instead of per cent predicted". In: *Breathe* 20.3. Review — Lung Function Corner, p. 230227. DOI: 10.1183/20734735.0227-2023. URL: https://publications.ersnet.org/content/breathe/20/3/230227.

Pandey, Anil Kumar et al. (Aug. 2021). "Fuzzy logic-based moving average filters for reducing noise from Tc-99m-sestamibi parathyroid images". In: *Nuclear Medicine Communications* 42.8, pp. 855–865. DOI: 10.1097/MNM.0000000000001410.

Perafan-Lopez, Juan Carlos et al. (June 2022). "Performance Analysis and Architecture of a Clustering Hybrid Algorithm Called FA+GA-DBSCAN Using Artificial Datasets". In: *Entropy (Basel)* 24.7, p. 875. DOI: 10.3390/e24070875.

Ponce, Mario C., Abdulghani Sankari, and Sandeep Sharma (2023). *Pulmonary Function Tests*. Last Update: August 28, 2023. StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/books/NBK482339/.

Quanjer, Philip H. et al. (2012). "Multi-ethnic reference values for spirometry for the 3–95 year age range: the Global Lung Function 2012 equations. Report of the Global Lung Function Initiative (GLI), ERS Task Force to establish improved Lung Function Reference

Values". In: *European Respiratory Journal* 40.6, pp. 1324–1343. DOI: 10.1183/09031936.00080312. URL: https://doi.org/10.1183/09031936.00080312.

Rohr, Maurice et al. (2024). "An extensive quantitative analysis of the effects of errors in beat-to-beat intervals on all commonly used HRV parameters". In: *Scientific Reports* 14. Open access, Published: 30 January 2024, p. 2498. DOI: 10.1038/s41598-023-50701-4. URL: https://www.nature.com/articles/s41598-023-50701-4.

Sadiya, S, T Alhanai, and M M Ghassemi (June 2021). "Artifact Detection and Correction in EEG data: A Review". In: *arXiv preprint arXiv:2106.04945*. Submitted on 10 Jun 2021. arXiv: 2106.04945.

Sakemi, Yusuke et al. (2024). "Learning reservoir dynamics with temporal self-modulation". In: *Communications Physics* 7, p. 29. DOI: 10.1038/s42005-024-00229-1. URL: https://www.nature.com/articles/s42005-024-00229-1.

Scano, Giorgio, Giulia Innocenti-Bruni, and Loredana Stendardi (2010). "Do obstructive and restrictive lung diseases share common underlying mechanisms of breathlessness?" In: *Respiratory Medicine* 104.7. Epub 2010 Mar 19, pp. 925–933. DOI: 10.1016/j.rmed.2010.02.019. URL: https://pubmed.ncbi.nlm.nih.gov/20303724.

Schermer, T.R. et al. (Oct. 2003). "Validity of spirometric testing in a general practice population of patients with chronic obstructive pulmonary disease (COPD)". In: *Thorax* 58.10, pp. 861–866. DOI: 10.1136/thorax.58.10.861.

Scrucca, Luca et al. (Aug. 2016). "mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models". In: *The R Journal* 8.1, pp. 289–317.

Singh, Garima et al. (2023). "Muco-Obstructive Lung Disease: A Systematic Review". In: *Cureus* 15.10, e46866. DOI: 10.7759/cureus.46866. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10637992/.

Stanojevic, Sanja, David A. Kaminsky, Martin R. Miller, et al. (2022). "ERS/ATS technical standard on interpretive strategies for routine lung function tests". In: *European Respiratory Journal* 60.1, p. 2101499. DOI: 10.1183/13993003.01499-2021. URL: https://doi.org/10.1183/13993003.01499-2021.

Tilert, Timothy et al. (Oct. 2013). "Estimating the US Prevalence of Chronic Obstructive Pulmonary Disease Using Pre- and Post-bronchodilator Spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007–2010". In: *Respiratory Research* 14.1, p. 103. DOI: 10.1186/1465-9921-14-103.

Vapnik, Vladimir N. (2010). *The Nature of Statistical Learning Theory*. New York: Springer.

Vermaak, J.C., A.E. Bunn, and M.A. de Kock (1979). "A new lung function index: The area under the maximum expiratory flow-volume curve". In: *Respiration* 37.2, pp. 61–65. DOI: 10.1159/000194008.

Verstraeten, D. et al. (2007). "An experimental unification of reservoir computing methods". In: *Neural Networks* 20.3, pp. 391–403. DOI: 10.1016/j.neunet.2007.04.003. URL: https://doi.org/10.1016/j.neunet.2007.04.003.

Wang, Yimin et al. (2022). "Deep learning for spirometry quality assurance with spirometric indices and curves". In: *Respiratory Research* 23, p. 98. DOI: 10.1186/s12931-022-02014-9. URL: https://doi.org/10.1186/s12931-022-02014-9.

Wang, Yongyu (2024). "Enabling DBSCAN for Very Large-Scale High-Dimensional Spaces". In: *arXiv preprint arXiv:2411.11421*. Version 3, last revised 3 Dec 2024. DOI: 10.48550/arXiv.2411.11421. URL: https://doi.org/10.48550/arXiv.2411.11421.

Wani, Aasim Ayaz (2024). "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions". In: *PeerJ Computer Science* 10. Published 29 Aug 2024, e2286. DOI: 10.7717/peerj-cs.2286. URL: https://doi.org/10.7717/peerj-cs.2286.

Wongoutong, Chantha (2024). "The impact of neglecting feature scaling in k-means clustering". In: *PLoS One* 19.12, e0310839. DOI: 10.1371/journal.pone.0310839.

Yamamoto, Shoichiro et al. (Oct. 2017). "Use of the forced-oscillation technique to estimate spirometry values". In: *International Journal of Chronic Obstructive Pulmonary Disease* 12, pp. 2859–2868. DOI: 10.2147/COPD.S143721.

Yang, Cheng-Hong et al. (2025). "An autoencoder-based arithmetic optimization clustering algorithm to enhance principal component analysis to study the relations between industrial market stock indices in real estate". In: *Expert Systems with Applications* 266. Published 25 March 2025, p. 126165. DOI: 10.1016/j.eswa.2024.126165. URL: https://doi.org/10.1016/j.eswa.2024.126165.

Yildiz, Izzet B., Herbert Jaeger, and Stefan J. Kiebel (2012). "Re-visiting the echo state property". In: *Neural Networks* 35. Epub 2012 Jul 23, pp. 1–9. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.07.005.

Yu, Liang et al. (Oct. 2011). "A degree-distribution based hierarchical agglomerative clustering algorithm for protein complexes identification". In: *Computational Biology and Chemistry* 35.5. Epub 2011 Jul 20, pp. 298–307. DOI: 10.1016/j.compbiolchem.2011.07.005.

Yuan, Nancy F. et al. (Dec. 2022). "Unsupervised learning identifies computed tomographic measurements as primary drivers of progression, exacerbation, and mortality in chronic obstructive pulmonary disease". In: *Annals of the American Thoracic Society* 19.12, pp. 1993–2002. DOI: 10.1513/AnnalsATS.202110-1127OC.

Zhao, Shitao et al. (Mar. 2018). "Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results". In: *Biological Procedures Online* 20.5. DOI: 10.1186/s12575-018-0067-8. URL: https://pubmed.ncbi.nlm.nih.gov/29507534/.

Zhou, Ruishi et al. (Mar. 2022). "Prediction of Pulmonary Function Parameters Based on a Combination Algorithm". In: *Bioengineering* 9.4, p. 136. DOI: 10.3390/bioengineering9040136. URL: https://www.mdpi.com/2306-5354/9/4/136.

Zhu, Ailin et al. (2021). "An Improved K-Means Algorithm Based on Evidence Distance". In: *Entropy* 23.11. Editors: Zoran H. Peric, Vlado Delic, Vladimir Despotovic, p. 1550. DOI: 10.3390/e23111550.