



PDF Download
3577190.3614109.pdf
01 February 2026
Total Citations: 2
Total Downloads: 814

Latest updates: <https://dl.acm.org/doi/10.1145/3577190.3614109>

RESEARCH-ARTICLE

Breathing New Life into COPD Assessment: Multisensory Home-monitoring for Predicting Severity

ZIXUAN XIAO, Swiss Federal Institute of Technology, Zurich, Zurich, ZH, Switzerland

MICHAL MUSZYNSKI, IBM Research - Zurich, Ruschlikon, Switzerland

RIČARDS MARCINKEVIČS, Swiss Federal Institute of Technology, Zurich, Zurich, ZH, Switzerland

LUKAS ZIMMERLI, IBM Research - Zurich, Ruschlikon, Switzerland

ADAM DANIEL IVANKAY, IBM Research - Zurich, Ruschlikon, Switzerland

DARIO KOHLBRENNER, University Hospital Zurich, Zurich, Switzerland

[View all](#)

Open Access Support provided by:

[IBM Research - Zurich](#)

[Swiss Federal Institute of Technology, Zurich](#)

[University Hospital Zurich](#)

Published: 09 October 2023

[Citation in BibTeX format](#)

ICMI '23: INTERNATIONAL
CONFERENCE ON MULTIMODAL
INTERACTION

October 9 - 13, 2023
Paris, France

Conference Sponsors:
[SIGCHI](#)

Breathing New Life into COPD Assessment: Multisensory Home-monitoring for Predicting Severity

Zixuan Xiao
ETH Zurich
Zurich, Switzerland
zixuanxiao1010@gmail.com

Lukas Zimmerli
IBM Research Europe
Zurich, Switzerland
lukas.zimmerli@gmail.com

Manuel Kuhn
University Hospital Zurich,
University of Zurich
Zurich, Switzerland
Manuel.Kuhn@usz.ch

Christian Clarenbach
University Hospital Zurich
Zurich, Switzerland
Christian.Clarenbach@usz.ch

Michal Muszynski
IBM Research Europe
Zurich, Switzerland
MMU@zurich.ibm.com

Adam Ivankay
IBM Research Europe
Zurich, Switzerland
AIV@zurich.ibm.com

Yves Nordmann
docdok.health
Basel, Switzerland
yves.nordmann@docdok.health

Julia E. Vogt
ETH Zurich
Zurich, Switzerland
julia.vogt@inf.ethz.ch

Ričards Marcinkevičs
ETH Zurich
Zurich, Switzerland
ricards.marcinkevics@inf.ethz.ch

Dario Kohlbrenner
University Hospital Zurich,
University of Zurich
Zurich, Switzerland
Dario.Kohlbrenner@usz.ch

Ulrich Muehlner
docdok.health
Basel, Switzerland
ulrich.muehlner@docdok.health

Thomas Brunschweiler
IBM Research Europe
Zurich, Switzerland
TBR@zurich.ibm.com

ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a significant public health issue, affecting more than 100 million people worldwide. Remote patient monitoring has shown great promise in the efficient management of patients with chronic diseases. This work presents the analysis of the data from a monitoring system developed to track COPD symptoms alongside patients' self-reports. In particular, we investigate the assessment of COPD severity using multisensory home-monitoring device data acquired from 30 patients over a period of three months. We describe a comprehensive data pre-processing and feature engineering pipeline for multimodal data from the remote home-monitoring of COPD patients. We develop and validate predictive models forecasting i) the absolute and ii) differenced COPD Assessment Test (CAT) scores based on the multisensory data. The best obtained models achieve Pearson's correlation coefficient of 0.93 and 0.37 for absolute and differenced CAT scores. In addition, we investigate the importance of individual sensor modalities for predicting CAT scores using group sparse regularization techniques. Our results suggest that feature groups indicative of the patient's general condition, such as static medical and physiological information, date, spirometer, and air quality,

are crucial for predicting the absolute CAT score. For predicting changes in CAT scores, sleep and physical activity features are most important, alongside the previous CAT score value. Our analysis demonstrates the potential of remote patient monitoring for COPD management and investigates which sensor modalities are most indicative of COPD severity as assessed by the CAT score. Our findings contribute to the development of effective and data-driven COPD management strategies.

CCS CONCEPTS

• **Applied computing** → **Health informatics; Health care information systems; Computing methodologies** → *Regularization; Neural networks.*

KEYWORDS

digital health; healthcare; remote patient monitoring; chronic obstructive pulmonary disease; grouped feature importance; predictive analysis; time series



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICMI '23, October 09–13, 2023, Paris, France
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0055-2/23/10.
<https://doi.org/10.1145/3577190.3614109>

ACM Reference Format:

Zixuan Xiao, Michal Muszynski, Ričards Marcinkevičs, Lukas Zimmerli, Adam Ivankay, Dario Kohlbrenner, Manuel Kuhn, Yves Nordmann, Ulrich Muehlner, Christian Clarenbach, Julia E. Vogt, and Thomas Brunschweiler. 2023. Breathing New Life into COPD Assessment: Multisensory Home-monitoring for Predicting Severity. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23)*, October 09–13, 2023, Paris, France. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3577190.3614109>

1 INTRODUCTION

Chronic obstructive pulmonary disease (COPD) refers to a group of conditions that cause airflow blockage and breathing-related problems, leading to progressive, irreversible limitations of airflow in the lungs [3]. COPD affects 174.5 million people worldwide (2.4% of the global population), according to data from 2015 [10]. It is a leading cause of morbidity and mortality, with a significant economic and social burden. An essential part of chronic disease management is regular assessment and monitoring of the disease progression and physical examination, which are labor-intensive.

The severity of COPD is sometimes assessed using patient's self-reports on physiological well-being—COPD Assessment Test (CAT) [23]. The CAT has the format of a questionnaire consisting of 8 questions about cough frequency, the quantity of sputum, chest tightness, breathlessness upon physical activities, limitation in doing home activities, confidence for going out, sleep quality and energy level. Each question needs to be answered with a score from 0 to 5, with higher scores indicating more severe COPD symptoms. However, the CAT requires active input from patients. Furthermore, the response is subjective and is provided during rare meetings with the physician [26].

There has been growing interest in using Remote Patient Monitoring (RPM) devices [22, 37, 38, 47] to facilitate chronic disease management at home. Firstly, RPM could be easily scaled up to incorporate more patients and health aspects into continuous monitoring since it is not limited by the presence or availability of healthcare workers. Secondly, RPM allows healthcare providers to detect health issues before they become severe based on the daily measurements of individuals' physiology and behaviors [25]. Thirdly, unlike subjective self-reports, sensor measurements provide an objective and quantitative assessment of the patient's state.

On top of the current progress in RPM for COPD, this study aims to evaluate the feasibility of using multisensory home-monitoring devices to assess COPD severity with the ultimate goal of improving disease management and patient treatment in the future. We specifically address two primary research questions (RQ):

RQ1: *Can we predict the COPD symptom severity indicated by the CAT score based on the past physiological and behavioral signals from multisensory home-monitoring devices?*

RQ2: *Which modalities and sensors are critically important for the CAT score prediction and, thus, COPD management?*

The contributions of this work are as follows. *i) Multimodal Data Fusion:* A pipeline is developed for pre-processing data from multiple physiological and behavioral signals collected using a home-monitoring device, CAir-desk [25] to obtain features to be used for the machine learning model development. *ii) CAT Prediction:* We explore different settings and methods for the predictive modeling of the CAT score. Obtained results may be used as guidelines and benchmarks in future studies on COPD patient home-monitoring. In particular, we formulate severity assessment as a prediction task and investigate predicting absolute values and changes in CAT scores from consecutive days. *iii) Digital Biomarker Importance:* We utilize different regularization techniques to investigate the importance of digital biomarker groups for CAT score prediction. As a technical contribution, we extend the neural-network-based

Granger causality framework [43] with feature grouping by incorporating group sparse regularization. Our findings may be valuable for improving COPD home-monitoring devices and future study design.

2 RELATED WORK

Remote Monitoring of COPD Patients. Several studies have investigated the home-monitoring of COPD patients [17, 37, 38, 47], as well as applying machine learning techniques to facilitate COPD management [4, 12]. For example, the mobile telehealth system *mHealth* [17] was designed to improve the self-management of COPD patients and detect acute exacerbations in a timely manner. Using this system, COPD patients were able to record their symptoms and medication use daily. Pulse rate and oxygen saturation were measured using a pulse oximeter and transmitted wirelessly to a tablet computer. This system required a 6-week calibration period to find personalized thresholds for alert signals. Around 40% of exacerbations were correctly detected three days before starting medication [17]. Another study [47] leveraged machine learning to predict exacerbation using data from wearables, including a home air quality sensor and a personal health advice mobile application. The resulting acute exacerbation of COPD (AECOPD) predictive model achieved an accuracy of 92.1%, sensitivity of 94%, and specificity of 90.4% for predicting events within the 7-day horizon. The model showed that the daily steps walked, stairs climbed, and distance moved were the most informative predictors. Moreover, the model achieved higher accuracy than simpler baselines relying only on the questionnaire data.

Despite these advancements, research on incorporating multisensory home-monitoring into respiratory disease mitigation and management is still scarce, and the devices used are mainly limited to smartphones [22]. Moreover, most research on predictive modeling has focused on the AECOPD event classification, which does not constitute a very fine-grained characterization of the patient's state. Furthermore, previous studies have required an asynchronous review of the daily monitoring data by a caregiver [38]. Thus, the design of a comprehensive, efficient and user-friendly home-monitoring system and the fine-grained assessment of COPD severity remain under-explored research questions.

Physiological and Behavioral Biomarkers. Due to the complexity of the disease, there is no exclusive biomarker for the assessment of COPD, but rather a combination of signs is used [36]. Some common physiological biomarkers include spirometry, exercise tolerance tests, and body composition analysis [39, 42]. Spirometry is part of the Global Initiative for Obstructive Lung Disease (GOLD) [1] standards for COPD assessment. Exercise tolerance tests, such as the six-minute walk test [39], can provide information on a patient's functional capacity and overall fitness level. Body composition analysis, including the body mass index (BMI), fat mass, and muscle mass measures, gives information on a patient's overall health status [42]. Behavioral biomarkers, such as smoking status, physical activity levels, and quality of life measures, help assess disease progression and treatment response in COPD. Smoking is the primary cause of COPD, and monitoring patient's smoking status is important for effective disease management. Physical activity levels and quality of life measures encapsulate the impact of COPD

on a patient’s daily life. They were used in several other studies as predictors of the AECOPD event [37, 47].

3 DATASET

During CAir study, we collected multisensory home-monitoring data from COPD patients. Our subjects were diagnosed with COPD according to GOLD guidelines, at least 40 years old, non-pregnant local language speakers without acute or recent exacerbation within the last 6 weeks and not attending a pulmonary rehabilitation program within the last 3 months [16]. They were home-monitored for around 12 weeks using CAir-desk [25]. CAir-desk (Figure 1a) is a novel multisensory home-monitoring device combining multiple sensors in a compact form with a single power plug for device charging. In particular, it includes a smartphone hub for collecting nocturnal cough recordings and sputum images, a Fitbit serving as a vital-sign and activity tracker, an air quality monitor, and a spirometer (Table 1). The participants were instructed to place CAir-desk in their bedrooms and perform daily measurements following the prescribed schedule (Figure 1b). In addition, the patients were required to fill out CAT questionnaires on their mobile phone, hosted by docdok.health platform, on a daily basis to self-report their symptoms. All data was securely transmitted to a cloud database through the Global System for Mobile Communications (GSM).

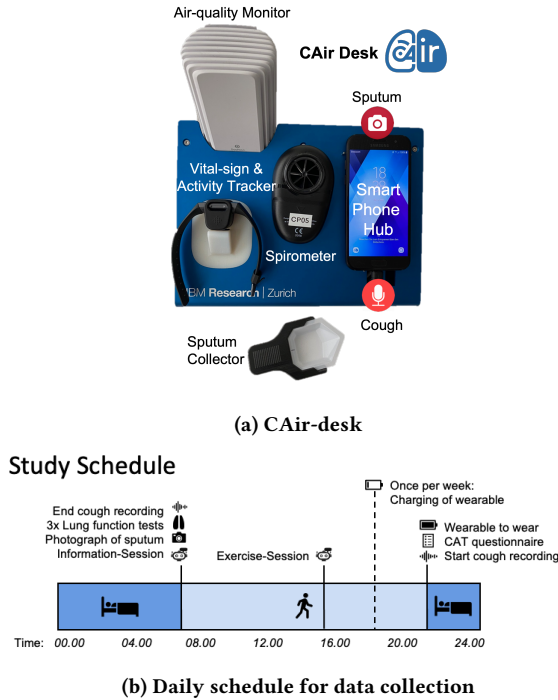


Figure 1: CAir study experimental setup. (a) CAir-desk, a novel multisensory device designed for home-monitoring COPD patients. (b) Data collection schedule prescribed to the study participants.

Multimodal data was acquired from 30 study participants. Their characteristics at the onboarding are reported in Table 2. The measurements came from several modalities described in detail in Table 1 alongside acquisition methods, quality control criteria applied to filter the data and patient adherence statistics.

4 METHODS

4.1 Multimodal Data Pre-processing and Feature Extraction

The data pre-processing and digital biomarker extraction pipeline is designed to be modality-specific. For each modality, it comprises five steps: *i) data retrieval*, *ii) data format cleaning and standardization*, *iii) feature extraction*, *iv) quality control (QC) filtering* and *v) alignment of resulting time series*. The feature extraction step is especially crucial for handling raw audio recording data. To extract nocturnal-cough-related features, we first apply the PANNs (pre-trained audio neural network) sound event detection and audio tagging model [27] to pre-processed truncated audio recordings. We use a probability threshold of 0.1 for at least 10 contiguous data points to select sound events labeled as *cough* and *throat clearing* with high certainty. Lastly, these events are aggregated by day as the cumulative event duration and count. Different QC filters are applied to each modality. In the raw data, there are 29 subjects with at least 1 entry of CAT score. As a quality control step, subjects with less than 10 days of CAT score records are discarded, resulting in a remainder of 23 subjects and 1523 days of data. Modality-specific and daily use adherence thresholds (Table 1) are applied to discard individual study days with high missingness rates.

4.2 Predictive Modeling of the CAT Score

To investigate the association between the multisensory recordings and COPD symptom severity given by the self-reported CAT score, we develop and validate (auto)regressive time series models forecasting the future CAT score using RPM data from the past. Below we provide details of the model development and evaluation.

4.2.1 Absolute and Differenced CAT Scores. We consider two different formats for representing the CAT score time series. The first one is the *absolute CAT score*, ranging from 0 to 40, henceforth referred to as *abs. CAT* (CAT_{abs}). The second one is the *differenced CAT score*, or *delta CAT score* (CAT_{Δ}), defined for each day t and each subject as

$$CAT_{\Delta}(t) = CAT_{abs}(t) - CAT_{abs}(t - 1). \quad (1)$$

Note that for the first observation for each subject, $CAT_{\Delta}(0)$ is undefined. In our analysis, we develop and validate time series models for predicting both $CAT_{abs}(t)$ and $CAT_{\Delta}(t)$ based on the past multisensory measurements represented as a multivariate time series $\{\mathbf{x}_t\}$. The two different response variables (absolute and differenced scores) allow exploring the use of home-monitoring data to quantify *i) the overall COPD impact on the patient and ii) day-to-day fluctuations in the patient’s state.*

4.2.2 Input Features. The predictors $\{\mathbf{x}_t\}$ consist of multiple modalities. After pre-processing (Section 4.1), all modalities are represented as tabular data comprising handcrafted physiological and behavioral markers. Therefore, modality-specific features are directly

Table 1: Summary of the data modalities, sources and features, alongside the quality control (QC) criteria, based on the domain knowledge, applied as inclusion conditions and the subjects' adherence to the study protocol, reported as median percentages and interquartile ranges.

Modality	Data Source	Features	# Features	QC Criteria	Adherence, %
Basic	—	Demographics, onboarding, offboarding and dropout dates and medical information	15	—	—
Date	—	Number of days since onboarding, whether the day falls into the COVID-19 lockdown and month	3	—	—
Air quality	Foobot, Airboxlab	Particulate matter, temperature and humidity, carbon dioxide in parts per million (CO ₂), volatile organic compounds, pollution score collected at minute resolution	6	≥ 20 hrs of data available per day	70 (27–95)
Activity	Charge 3, Fitbit	Number of calories burned, steps, distance, floors, minutes sedentary, minutes lightly active, minutes fairly active, minutes very active and calories burned during activities	9	≥ 20 hrs of data available or ≤ 20 minutes of data unavailable between 07:00 to 22:00 per day	47 (7–79)
Sleep	Charge 3, Fitbit	Number of minutes asleep, minutes awake, number of awakenings, time in bed, minutes of the Rapid Eye Movement (REM) sleep, minutes of light sleep and minutes of deep sleep	7	≥ 20 hrs of data available or ≥ 5 hrs 40 min of data available between 20:00 and 10:00 (+1) per day	41 (7–83)
Cough	Galaxy A320, Samsung	Count and duration of cough and throat clearing events	4	Recording duration between 4 and 10 hrs and the start time of recording between 19:00 and 02:00 (+1)	28 (5–63)
Spirometer	Air Next Spirometer, NuvoAir	Forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC) and FEV1/FVC	8	≥ 3 spirometer tests per day	18 (8–81)
Sputum	Galaxy A320, Samsung	Images of sputum in the sputum collector	—	Excluded from the analysis	0 (0–4)
CAT Score	CAT questionnaire, docdok.health	8 questions about the quality of life	8	≤ 3 days between expected and actual response dates	86 (28–97)

Table 2: Study participant characteristics at the onboarding. Continuous variables are reported as medians and interquartile ranges (IQR); categorically-valued covariates are reported as frequencies and percentages for the categories.

Characteristic	Statistics
Age, median (IQR), years	65 (59–70)
Self-identified gender (female/male), <i>n</i> (%)	13/17 (43%/57%)
BMI, median (IQR), kg/m ²	25.8 (22.0–29.0)
GOLD [1] criteria, (1/2/3/4), <i>n</i> (%)	3/14/9/4 (10%/47%/30%/13%)
Smoking status (yes/no), <i>n</i> (%)	7/23 (23%/77%)
Observation length, median (IQR), days	92 (65–99)

concatenated after down-sampling to daily resolution and temporal alignment. In addition, static patient information recorded at the onboarding is appended at every time step. Date-related features are also added to capture potential seasonality. In particular, we include the number of days since the onboarding, month and whether the date falls into the COVID-19 lockdown period in the corresponding location. For the differenced CAT score (Equation 1), we fit an *autoregressive* model with the previous day's score, $CAT(t-1)$, as a predictor. In this case, the previous CAT score is given by eight components corresponding to the individual questions.

Not all the models we train and validate (Section 4.2.3) account for the sequential nature of the data. Therefore, we restructure the inputs so that temporal dependencies can be captured by non-sequential predictive models. Thus, when predicting CAT from day t , we concatenate the inputs from t with the features from every

day between $t-1$ to $t-k$, where k denotes the maximum lag of (auto)regressive relationships. Considering clinicians' suggestions, the risk of overfitting on a small cohort and data loss associated with larger lags, in our analysis, we set $k = 3$ days.

Due to the patients' non-adherence and voluntary recording, some time series segments are missing. We impute the missing data before training the models using multivariate imputation by chained equations [45], an imputation strategy based on modeling missing values as a function of other features.

4.2.3 Predictive Model Development and Validation. To comprehensively evaluate and explore the relationship between the multisensory home-monitoring data and CAT score, we fit linear and non-linear models with the above inputs and responses. We experiment with linear regression models with Lasso [44] and Ridge regularization [20], tree-based ensembles, including Random Forests (RF) [5] and Extreme Gradient Boosting (XGB) [8], and neural networks (NN), specifically multilayer perceptrons (MLP) and Long Short-Term Memory (LSTM) [19]. Additionally, to explore the importance of modalities, we consider regularized neural nets, such as component-wise LSTM [43] and LassoNet [30] (Section 4.3). All models are trained by minimizing the mean squared error (MSE) loss.

Our dataset comprises multiple time series replicates, each acquired from a different subject. We split the data into the training, test and holdout sets. The holdout set consists of the *whole* time series from 5–6 subjects. We randomly split our cohort of 23 subjects into four groups of 5–6 subjects each. We repeated the process

with different random seeds four times (random permutation cross-validation blocked by the subject), generating 16 holdout sets. The data from subjects not in the holdout set is split into five folds blocked by time [40]. The first four folds are used for training and hyperparameter tuning, whereas the last fold is used for testing.

Due to the imbalance in the distribution of CAT score values (Section 5.1), we choose macro-averaged mean absolute error (MAMAE) [31, 46] as the model selection criterion during cross-validation. For the test and holdout sets, we report predictive performance in terms of Pearson's correlation coefficient, mean absolute error (MAE) and MAMAE. Since there are no state-of-the-art results on CAT score prediction from multisensory data to compare against, we include two simple baseline models for reference. The first baseline is the *naive model* that uses the CAT score from the previous day, $CAT(t-1)$, as the prediction for the day t . Since the CAT time series is highly autocorrelated (Section 5.1), it is essential to benchmark the models' performance against the naive baseline. The second reference is the *constant model* that always outputs the mean CAT value of the corresponding subject as a prediction. Since the distribution of the CAT scores is highly imbalanced (Figure 2), we expect the constant model to perform relatively well.

4.3 Group Sparse Regularization and Feature Group Importance

We utilize models with structured sparsity regularization [50] to investigate the contributions of individual features and feature groups defined by modalities (Table 1) to the forecast of the CAT score. Such models implicitly perform feature selection and are, thus, deemed more parsimonious and interpretable [33]. Below we briefly explain the sparse regularization approaches we applied and describe their extension to grouped features.

4.3.1 Linear Group Lasso. In addition to linear models with Lasso and Ridge regularization, we consider *group Lasso* regularization [34, 50]. In group Lasso, the ℓ_1 penalty is applied to the groups of coefficients; and in our context, these groups are defined by the different data modalities. Consider a linear regression problem with $\beta = (\beta_1, \beta_2, \dots, \beta_p)^\top$ being p regression parameters that can be arranged into m groups as $\beta_{\mathcal{G}} = \{\beta_{\mathcal{G}_1}, \beta_{\mathcal{G}_2}, \dots, \beta_{\mathcal{G}_m}\}$, where for $1 \leq j \leq m$, $\mathcal{G}_j = \{j_1, \dots, j_{p_j}\} \subseteq \{1, \dots, p\}$ is the set of indices belonging to the j -th group of size p_j and $\beta_{\mathcal{G}_j}$ are weights indexed by \mathcal{G}_j . Henceforth, we assume that the grouping is i) non-overlapping, i.e., $\mathcal{G}_j \cap \mathcal{G}_k = \emptyset$, for all $j \neq k$, and ii) totally covering, i.e., $\bigcup_{j=1}^m \mathcal{G}_j = \{1, \dots, p\}$ [13]. The optimization problem for the linear regression with the group Lasso regularization is given by [13, 34]

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^m \sqrt{p_j} \|\beta_{\mathcal{G}_j}\|_2, \quad (2)$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, n is the number of samples, and $\lambda > 0$ is the weight of the penalty term. In our analysis, the design matrix X is constructed using multiple lagged values from each predictor time series, y is given by the future absolute or differenced CAT score, and groups $1 \leq j \leq m$ are defined by modalities (Section 4.2).

4.3.2 Sparse-input Neural Networks. The group Lasso penalty term from Equation 2 can be readily generalized to neural network models [11, 30, 41, 43] to induce structured sparsity in the input layer and the whole architecture.

In the time series context, sparse neural networks [14, 24, 32, 43] have been extensively used to infer Granger causality [15], i.e., to infer time series nonlinear autocorrelation structure. For instance, component-wise MLPs (cMLP) and LSTMs (cLSTM) proposed in [43] infer Granger causes of the response time series y_t by minimizing the following loss:

$$\min_{\mathbf{W}} \sum_{t=k}^T \left(y_t - g \left(\mathbf{x}_{(t-k):(t-1)} \right) \right)^2 + \lambda \sum_{j=1}^p \|\mathbf{W}_{:j}^1\|_F, \quad (3)$$

where k is the maximum lag of autoregressive relationships, T is the length of the observed time series, $g(\cdot)$ is a neural network parameterized by weight matrices \mathbf{W} , \mathbf{W}^1 is the input-layer weight matrix, and $\mathbf{W}_{:j}^1 = (\mathbf{W}_{:j}^{11}, \dots, \mathbf{W}_{:j}^{1k})$ corresponds to the input weights of the j -th predictor time series across all lags.

LassoNet [30] implements a similar regularization scheme for static data. However, in addition to regularizing input-layer weights, it introduces a linear skip layer and a hierarchical sparsity-inducing penalty [9, 48]. The modified constrained objective is given by

$$\min_{\theta, \mathbf{W}} \mathcal{L}(\theta, \mathbf{W}) + \lambda \|\theta\|_1, \quad (4)$$

subject to $\|\mathbf{W}_{:j}^1\|_{\infty} \leq M |\theta_j|, j = 1, \dots, p,$

where $\mathcal{L}(\theta, \mathbf{W})$ is the empirical loss on the training set, e.g., MSE or cross-entropy, θ are the linear skip layer weights, $\mathbf{W}_{:j}^1$ are the weights of the j -th feature in the first nonlinear layer, and $\lambda, M > 0$ are hyperparameters determining the penalty weight. Similar to the group Lasso linear regression (Equation 2), LassoNet can be readily applied to time series data by introducing lagged time series values as separate features.

4.3.3 Neural Granger Causality for Grouped Features. Note that the penalty term in Equation 3 only groups lagged values from the *same* predictor time series [43]. We modify this penalty to incorporate feature groups, e.g., to group features originating from the same data modality (Table 1). Following the notation from Equations 2–4, let the input-layer weights \mathbf{W}^1 be arranged into m groups given by $\mathbf{W}_{\mathcal{G}}^1 = \{\mathbf{W}_{\mathcal{G}_1}^1, \dots, \mathbf{W}_{\mathcal{G}_m}^1\}$. Here, $\mathbf{W}_{\mathcal{G}_j}^1$, for $1 \leq j \leq m$, represents the weights belonging to the j -th group. Namely, $\mathbf{W}_{\mathcal{G}_j}^1 = (\mathbf{W}_{:j_1}^{11}, \dots, \mathbf{W}_{:j_1}^{1k}, \dots, \mathbf{W}_{:j_{p_j}}^{11}, \dots, \mathbf{W}_{:j_{p_j}}^{1k})$, where $\{j_1, \dots, j_{p_j}\} \subseteq \{1, \dots, p\}$ is the set of indices of the predictors from the j -th group. Thus, to incorporate grouped features, Equation 3 can be rewritten as

$$\min_{\mathbf{W}} \sum_{t=k}^T \left(y_t - g \left(\mathbf{x}_{(t-k):(t-1)} \right) \right)^2 + \lambda \sum_{j=1}^m \|\mathbf{W}_{\mathcal{G}_j}^1\|_F. \quad (5)$$

In the same vein, the LassoNet objective (Equation 4) can be modified to impose a penalty on groups rather than single predictors.

4.3.4 Quantifying Feature Group Importance. To understand the contribution of each modality to the CAT score prediction, we apply linear group Lasso regression, cLSTM and LassoNet with group penalties (Section 4.3.3). We train linear and cLSTM models under

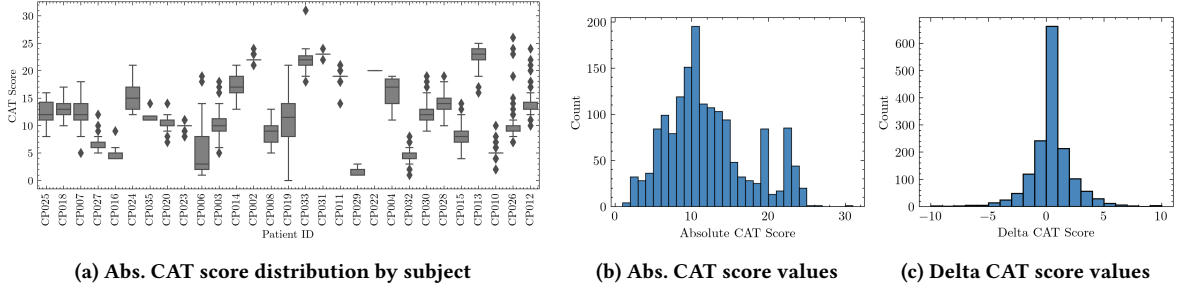


Figure 2: Absolute and differenced CAT score values and their distributions. (a) The box plot of reported CAT score values for each subject across the observation period. (b) The histogram of the absolute CAT score values, aggregated across the whole cohort. (c) The histogram of the differenced CAT score values, aggregated across the whole cohort.

varying group penalty weight λ values, incrementing from 0 with a step size of 0.001. For the LassoNet, we start with a heuristically-inferred value [30] and increment it by a path multiplier of 1.02. We increase λ until the weights of all feature groups are shrunk to zero. For the linear group Lasso, we define the weight of the group j as $\|\beta_{\mathcal{G}_j}\|_2$ (Equation 2). For the cLSTM and LassoNet, we use the Frobenius norm of the group weight matrix given by $\|\mathbf{W}_{\mathcal{G}_j}^1\|_F$. Consequently, the importance of the group is given by λ at which the group weight vanishes.

5 RESULTS

5.1 Exploring CAT Score Time Series

Before exploring the association between the multisensory home-monitoring data and CAT scores, we comprehensively characterize the CAT score time series to better understand patterns in the self-reported disease impact.

CAT Score Distribution. The distributions of CAT scores for each subject are shown in Figure 2a as box plots. In general, the within-subject variance of the CAT scores is small. On the other hand, the mean CAT scores for individual subjects range between 2 and 23. Within the cohort, absolute CAT score values are not evenly distributed across the value range (Figures 2a and 2b). Most of the absolute CAT score values fall into the range of 0–15; we observed few scores in the 25–31 range and no values from the 32–40 range. These results indicate that the patients from the CAir cohort have relatively mild symptoms. Differenced CAT score values mainly fall into the range of $[-2, 2]$, with a considerable percentage equal to 0 (Figure 2c), showing that the CAT score is temporally stable for individual subjects.

Stationarity and Autocorrelation. We apply the KPSS test [28] with a lag of 7 days and constant-trend stationarity assumption to the CAT score time series (at a significance level $\alpha = 0.10$). For the absolute CAT score for 8/23 patients, the null hypothesis of stationarity is rejected with at least 90% confidence. In contrast, for the differenced time series, the null hypothesis is rejected only in 2/23 subjects. Absolute CAT score shows a positive autocorrelation with a median of 0.41 (IQR: $[0.24, 0.64]$) for the lag of $k = 1$. Differenced CAT score features a weaker *negative* autocorrelation with a median of -0.33 (IQR: $[-0.40, -0.24]$) for lag $k = 1$.

5.2 COPD Severity Prediction

5.2.1 Absolute CAT Score Prediction. We first investigate the absolute CAT score prediction. The top panel of Figure 3a depicts the results of the test-set evaluation (blocked temporally). On the one hand, we observe that all models outperform the naive baseline (Section 4.2) w.r.t. Pearson’s correlation coefficient. In terms of the MAE and MAMAE, tree ensembles, followed by neural networks, have a lower error than the naive model. On the other hand, linear models tend to have higher MAE and MAMAE. We attribute these differences in performance to the nonlinear relationship between sensor inputs and the absolute CAT score. In addition, neural networks and tree-based ensembles benefit from subject-specific calibration.

On the holdout set (Figure 3a, bottom), none of the models outperforms the constant baseline w.r.t. all three metrics. We hypothesize that to predict the absolute CAT score accurately, models require information about the *baseline condition* of a subject. For example, it can be represented by the patient’s mean CAT score over a relatively long and stable period of time. We validate this hypothesis by considering the *delta* CAT score prediction on a holdout set (Section 5.2.2), where the need for knowing the baseline condition is removed by differencing. We also observe considerable variance across the holdout sets. Thus, models for the absolute CAT score prediction do not generalize to all new subjects.

We conduct further experiments to support our hypothesis about the role of the baseline condition. Specifically, we introduce the mean CAT score for each patient over the whole observation period and the CAT score on the first day of the study as input features. Figure 4 shows the results of this ablation. Using the mean CAT as an input makes holdout-set performance comparable to that on the test set. Incorporating the CAT score at onboarding yields a more minor yet visible improvement. In practice, it would be feasible to input the CAT score on the first day, as assessed during a meeting with the physician.

5.2.2 Differenced CAT Score Prediction. In contrast to the absolute CAT score, the delta CAT score can be predicted equally well on the test and holdout sets. Most models outperform the constant and naive baselines by a considerable margin in terms of Pearson’s correlation coefficient and MAMAE. However, the constant model, which predicts the mean CAT difference across the training cohort,

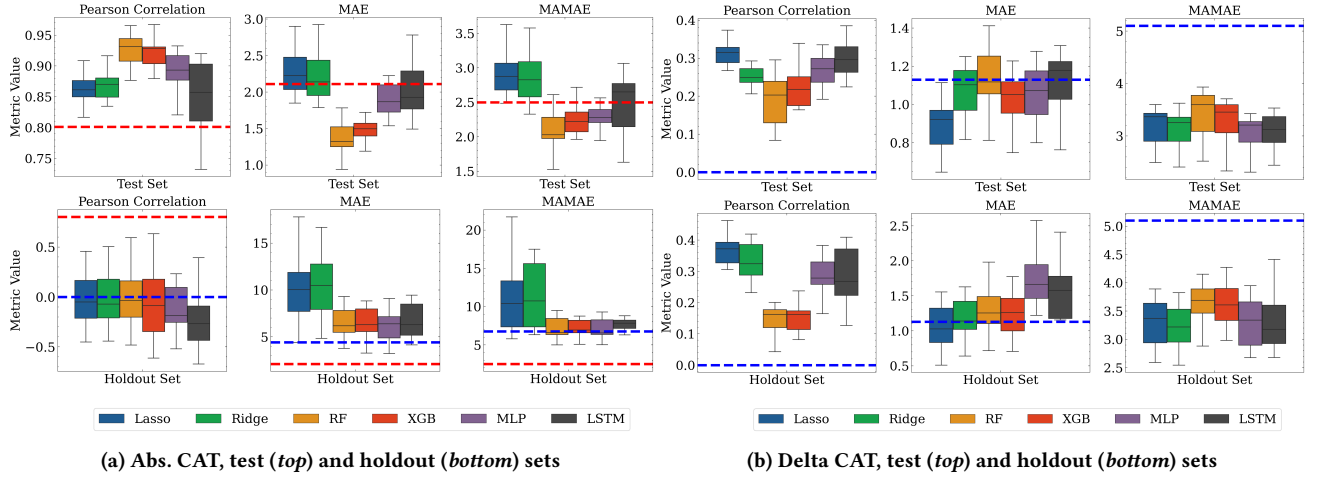


Figure 3: Test- and holdout-set results for the (a) absolute and (b) differenced CAT score prediction w.r.t. Pearson’s correlation coefficient (\uparrow), MAE (\downarrow) and MAMAE (\downarrow). Dashed red and blue lines correspond to the naive and constant baselines, respectively.

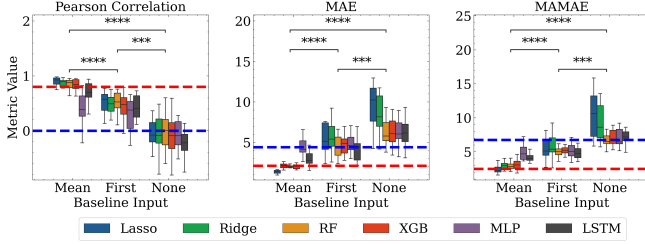


Figure 4: Holdout-set results for the absolute CAT score prediction with and without the baseline input. We tested for the differences in the performance of the models using the Wilcoxon signed-rank test. p -values: *, $0.0001 < p \leq 0.001$; ****, $p \leq 0.0001$.**

achieves a low MAE and is challenging to outperform. We attribute this to most CAT score differences being close to 0–1 (Figure 2c). Among the trained models, Lasso and Ridge regression perform better than the rest overall, suggesting that, expectedly, in our moderately sized cohort, regularization helps prevent overfitting.

5.2.3 Further Remarks. In summary, we analyze the CAT score time series as a composition of the baseline score that is relatively stable over time and day-to-day fluctuations. For predicting the differenced CAT score, i.e., daily changes, models based on sensor inputs and static information generalize well on unseen data, blocked temporally and by subject. This result suggests that CAT dynamics are likely, not subject-dependent. However, models for the absolute CAT score prediction generalize poorly on the holdout set. By introducing the patient’s mean CAT score as an input feature, we are able to reduce the discrepancy between test- and holdout-set performance. These results suggest that learning subject-dependent relationships from a cohort of only 23 patients is challenging.

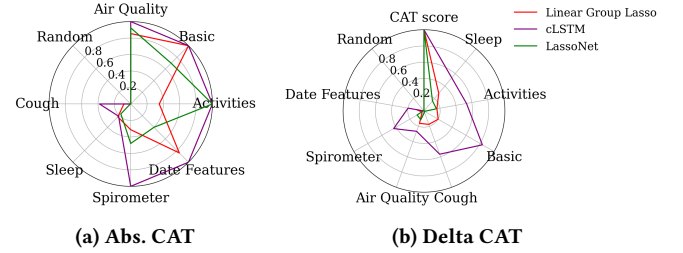


Figure 5: Summary of the feature group importance from the linear group Lasso, cLSTM, and LassoNet models for the (a) absolute and (b) differenced CAT score. Feature groups are ordered clockwise by the importance rank averaged across the methods. Importance values were scaled by min-max normalization.

5.3 Important Modalities

As explained in Section 4.3, we utilize group sparse regularization to explore the importance of individual modalities (Table 1). The feature group importances derived from the group Lasso linear regression, LassoNet and cLSTM models are to some degree consistent (Figure 5). For the absolute CAT score, the agreement among models is high, with a clear distinction between important feature groups, such as air quality, basic information, activities, date features, and spirometer, and modalities with only a minor impact, such as sleep and cough (Figures 5a). Features from the basic information and spirometry are likely to be associated with the patient’s general condition and are, thus, useful for predicting absolute CAT values. A plausible explanation for the high importance of air quality and date-related features is their ability to capture time-dependent trends, e.g., related to seasonal changes: we have shown before that the absolute CAT score time series is non-stationary for almost half of the subjects (Section 5.1).

Conversely, when predicting the differenced time series, the past CAT score is the top-ranked input feature (Figures 5b), followed

by the sleep and activity modalities. The date, spirometer and air quality groups are less important in this case. The role of the basic and cough features when predicting the delta CAT score remains unclear due to the large variation across the methods; namely, while cLSTM and linear group Lasso assign higher importance to these groups, LassoNet appears to ignore both.

For all three regularized models, we introduce a dummy ‘modality’ with a vector of three features per day drawn independently of the CAT score from the uniform distribution on $[0, 1]$ (‘Random’ in Figure 5). Assuringly, these random features are consistently ranked as the least important. Thus, according to this sanity check, none of the models infers trivial false positive associations.

Interestingly, the cough feature group is important for neither the absolute nor delta CAT score prediction, despite cough being among the primary symptoms of COPD. There exist a few plausible explanations for this finding. Firstly, the cough modality has the lowest adherence rate (Table 1), with a median adherence rate of 28% across subjects and approximately 50% of the data missing. The high missingness rate potentially discounts the reliability of the extracted features. Secondly, the feature extraction from the cough recordings has a relatively complex pipeline prone to introducing errors and additional unwanted variability. Last but not least, cough recordings were captured regardless of the subject’s presence in the room. Thus, additional feature noise could have been introduced by other sources of cough, e.g., other persons from the participants’ households.

6 DISCUSSION

In this study, we aimed to predict the severity of COPD disease, represented by the CAT score, through digital biomarkers extracted from multisensory home-monitoring data. To this end, we designed a specially tailored multisensory CAir-desk device for COPD patient monitoring and developed a data pre-processing pipeline comprising imputation, quality control and feature extraction steps for the downstream machine learning model development and time series analysis. Our study on RPM for COPD stands out from the previous research, as reviewed in [38], in several ways. Firstly, our data come from a more comprehensive set of sensors, providing a more holistic approach to monitoring. Additionally, unlike most of the work in this field that focuses solely on predicting acute exacerbation events [4, 37, 38, 47], we have considered a different target—the CAT score, which offers a finer-grained assessment of COPD impact on the patient.

To profile the patient cohort and better understand the dynamics of COPD severity over time, an exploratory data analysis was conducted. Our results suggest that the absolute CAT score changes slowly through time and often features a non-constant trend. The changes between consecutive days appear to be stationary for almost all subjects over a period of around three months. To address **RQ1** (Section 1), we investigated the predictive modeling of the CAT score time series based on multisensory measurements. Firstly, the models are able to forecast the absolute CAT score accurately without the past CAT results as an input, provided that the model could be calibrated on the data from the same subject. This finding was established by validating predictive models on the temporally-blocked test set. Thus, it is, in principle, possible to assess COPD

impact in an automated manner based on the home-monitoring data. Ultimately, multisensory devices similar to the one presented in this work may reduce the burden of tedious and subjective daily self-reports from the patient. Secondly, the changes in the CAT score can be modeled given the past CAT outcomes, with or *without* a patient-specific calibration period, as confirmed by our evaluation on both test- and holdout-set data. Consequently, it may be possible to detect clinically significant deterioration events based on the differenced CAT score predictions.

To address **RQ2**, we analyzed the importance of feature groups from different modalities for predicting the absolute and differenced CAT scores. To this end, we leveraged linear group Lasso regression, component-wise LSTMs and LassoNet models. For neural-network-based Granger-causal inference [43], we expanded the cMLP and cLSTM models to include feature groups defined, in addition to different lags, by modalities. We observed that static information, air quality, date and activity features and spirometry are essential for the absolute CAT score prediction. The importance of these features might be attributed to their utility in predicting the patient’s general condition and seasonal trend in CAT. Interestingly, sleep and cough are the least important modalities.

By contrast, differenced CAT score prediction relies heavily on the past CAT score as an input and sleep and activity features. The importance of activities and self-reported symptoms was also observed in another COPD RPM study investigating acute exacerbation event prediction [47]. Unlike for absolute CAT score prediction, air quality, spirometry and date features do not show prominent and consistent importance. This finding is also supported by another study aiming to identify signs of exacerbations in COPD patients [21], which suggests that spirometry results, although varying within a noticeable range, show weak or absent correlation with other prognostic COPD markers.

Our findings provide empirical support for adjusting and improving multisensory measurement setups in future studies. For example, depending on the goal of predicting either the patient’s general state or fluctuations, it may be necessary to focus on the different sets of sensors.

6.1 Limitations

There are several significant limitations in the current design of the study. The CAT score is merely a surrogate for the severity of COPD since it relies on self-reported data subject to personal interpretation. In addition, the moderately-sized cohort comprising only 30 patients observed for only three months does not provide a comprehensive representation of the subject-specific relationships and variance in the CAT score dynamics and introduces the risk of overfitting. Moreover, the current cohort only contains subjects with mild symptoms, not experiencing exacerbation events. Therefore, our findings and models may not transfer to the general population of COPD patients. Last but not least, the dataset was acquired from a single geographical area, and hence, our findings should be validated externally in a multicenter study.

6.2 Future Work

As a natural extension of this work, future research could focus on expanding the dataset to encompass a broader population of

patients with varying degrees of COPD severity and observe subjects for a longer period. In addition, the cough feature quality can be improved by attaining higher adherence rates, utilizing sex- or voice-specific audio tagging models [2] for feature extraction or relying on alternative data sources, e.g., accelerometry [35], to measure cough. From the modeling perspective, a potential improvement in predictive performance could be attained by personalized subgroup-level models trained on patient sub-populations, e.g., defined based on well-established medical criteria [1] or in a data-driven manner using unsupervised learning. In the current analysis, we considered only the task of one-day-ahead forecasting; however, longer-range prediction horizons might be more practical and should be investigated. Moreover, we approached the prediction problem exclusively from the perspective of time series regression. Alternative approaches, for instance, classifying clinically relevant deterioration or treating the CAT score as an ordinal variable, could provide additional insights. Last but not least, we observed that the distributions of the absolute and differenced CAT values were very concentrated; hence, a more systematic investigation of the imbalanced learning methods [6, 7, 18, 29, 49] could help attain better predictive performance.

7 CONCLUSION

This work shows that COPD severity can be forecast effectively using digital biomarkers recorded by the specially-tailored home-monitoring multisensory device [ANONYMIZED]-desk. We established a data pre-processing and feature engineering pipeline for the CAir-desk for the downstream machine learning model development and validation. The pipeline is compatible with other home-monitoring frameworks capturing similar physiological features. We trained and evaluated time series models in two different settings: predicting the absolute and differenced COPD assessment test scores. Our findings indicate that the absolute CAT scores can be predicted with high accuracy given a calibration period on the data from the same subject. A baseline input, such as the patient's CAT score at entry, is necessary to generalize the absolute CAT score prediction models to unseen subjects. Furthermore, the changes in CAT scores can be modeled given the past CAT score values with or without patient-specific calibration.

Our analysis also explored the importance of different modalities in predicting absolute and differenced CAT scores. Feature groups indicative of the general condition and trends, such as static information acquired at onboarding, date features and air quality, are crucial for the absolute CAT score prediction. For predicting fluctuations in the CAT score, sleep and activity features, reflecting daily symptomatic changes, are more important than air quality, spirometry and date features. These findings suggest that modalities should be carefully chosen and adjusted in future studies to capture COPD severity more comprehensively, reduce equipment costs, increase user adherence and provide further empirical support for COPD home-monitoring device development. More generally, this study contributes to the advancement of multisensory home-based monitoring for personalized COPD severity assessment, highlighting the potential of digital biomarkers to improve COPD monitoring and, consequently, management.

ACKNOWLEDGMENTS

We thank all participants to have participated in the CAir trial. This study was funded by Innosuisse project 29844.1. RM was supported by the Swiss National Science Foundation (SNSF) grant #320038189096.

REFERENCES

- [1] Alvar Agustí, Bartolome R. Celli, Gerard J. Criner, David Halpin, Antonio Anzueto, Peter Barnes, Jean Bourbeau, MeiLan K. Han, Fernando J. Martinez, Maria Montes de Oca, Kevin Mortimer, Alberto Papi, Ian Pavord, Nicolas Roche, Sundee Salvi, Don D. Sin, Dave Singh, Robert Stockley, M. Victorina López Varela, Jadwiga A. Wedzicha, and Claus F. Vogelmeier. 2023. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary. *The European Respiratory Journal* 61, 4 (April 2023), 2300239. <https://doi.org/10.1183/13993003.00239-2023>
- [2] Filipe Barata, Peter Tinschert, Frank Rassouli, Claudia Steurer-Stey, Elgar Fleisch, Milo Alan Puhon, Martin Brutsche, David Kotz, and Tobias Kowatsch. 2020. Automatic Recognition, Segmentation, and Sex Assignment of Nocturnal Asthmatic Coughs and Cough Epochs in Smartphone Audio Recordings: Observational Field Study. *Journal of Medical Internet Research* 22, 7 (2020), e18082. <https://doi.org/10.2196/18082>
- [3] Peter J. Barnes, Peter G. J. Burney, Edwin K. Silverman, Bartolome R. Celli, Jørgen Vestbo, Jadwiga A. Wedzicha, and Emiel F. M. Wouters. 2015. Chronic obstructive pulmonary disease. *Nature Reviews Disease Primers* 1, 1 (2015), 15076. <https://doi.org/10.1038/nrdp.2015.76>
- [4] Lemana Spahić Bećirović, Amar Deumić, Lejla Gurbeta Pokvić, and Almir Badnjević. 2021. Artificial Intelligence Challenges in COPD management: a review. In *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, Kragujevac, Serbia, 1–7. <https://doi.org/10.1109/BIBE52308.2021.9635374>
- [5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [6] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research* 16, 1 (2002), 321–357.
- [7] Chao Chen and Leo Breiman. 2004. *Using Random Forest to Learn Imbalanced Data*. Technical Report. Department of Statistics, University of California, Berkeley.
- [8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [9] Nam Hee Choi, William Li, and Ji Zhu. 2010. Variable Selection With the Strong Heredity Constraint and Its Oracle Property. *J. Amer. Statist. Assoc.* 105, 489 (2010), 354–364. <https://doi.org/10.1198/jasa.2010.tm08281>
- [10] GBD 2015 Chronic Respiratory Disease Collaborators. 2017. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Respiratory Medicine* 5, 9 (2017), 691–706. [https://doi.org/10.1016/s2213-2600\(17\)30293-x](https://doi.org/10.1016/s2213-2600(17)30293-x)
- [11] Jean Feng and Noah Simon. 2022. Ensembled sparse-input hierarchical networks for high-dimensional datasets. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15, 6 (2022), 736–750. <https://doi.org/10.1002/sam.11579>
- [12] Yinhe Feng, Yubin Wang, Chunfang Zeng, and Hui Mao. 2021. Artificial Intelligence and Machine Learning in Chronic Airway Diseases: Focus on Asthma and Chronic Obstructive Pulmonary Disease. *International Journal of Medical Sciences* 18, 13 (2021), 2871–2889. <https://doi.org/10.7150/ijms.58191>
- [13] Jordan Frecon, Saverio Salzo, and Massimiliano Pontil. 2018. Bilevel Learning of the Group Lasso Structure. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) (NIPS '18). Curran Associates Inc., Red Hook, NY, USA, 8311–8321.
- [14] Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, and Jingping Bi. 2023. Causal Discovery from Temporal Data: An Overview and New Perspectives. *arXiv:2303.10112*
- [15] Clive W. J. Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37, 3 (1969), 424. <https://doi.org/10.2307/1912791>
- [16] Christoph Gross, Dario Kohlbrenner, Christian F. Clarenbach, Adam Ivankay, Thomas Brunschweiler, Yves Nordmann, and Florian V Wangenheim. 2020. A Telemonitoring and Hybrid Virtual Coaching Solution "CAir" for Patients with Chronic Obstructive Pulmonary Disease: Protocol for a Randomized Controlled Trial. *JMIR research protocols* 9, 10 (Oct. 2020), e20412. <https://doi.org/10.2196/20412>

- [17] Maxine Hardinge, Heather Rutter, Carmelo Velardo, Syed Ahmar Shah, Veronika Williams, Lionel Tarassenko, and Andrew Farmer. 2015. Using a mobile health application to support self-management in chronic obstructive pulmonary disease: a six-month cohort study. *BMC Medical Informatics and Decision Making* 15, 1 (2015). <https://doi.org/10.1186/s12911-015-0171-5>
- [18] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, Hong Kong, China, 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
- [19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [20] Arthur E. Hoerl and Robert W. Kennard. 2000. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 42, 1 (2000), 80–86. <http://www.jstor.org/stable/1271436>
- [21] Åsa Holmner, Fredrik Ohlberg, Urban Wiklund, Eva Bergmann, Anders Blomberg, and Karin Wadell. 2020. How stable is lung function in patients with stable chronic obstructive pulmonary disease when monitored using a telehealth system? A longitudinal and home-based study. *BMC Medical Informatics and Decision Making* 20 (2020). <https://doi.org/10.1186/s12911-020-1103-6>
- [22] Sadia Janjua, Emma Banchoff, Christopher J.D. Threapleton, Samantha Prigmore, Joshua Fletcher, and Rebecca T. Disler. 2021. Digital interventions for the management of chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews* 2021, 4 (2021). <https://doi.org/10.1002/14651858.cd013246.pub2>
- [23] P. W. Jones, G. Harding, P. Berry, I. Wiklund, W-H. Chen, and N. Kline Leidy. 2009. Development and first validation of the COPD Assessment Test. *European Respiratory Journal* 34, 3 (2009), 648–654. <https://doi.org/10.1183/09031936.00102509>
- [24] Saurabh Khanna and Vincent Y. F. Tan. 2020. Economy Statistical Recurrent Units For Inferring Nonlinear Granger Causality. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, Addis Ababa, Ethiopia. <https://openreview.net/forum?id=SyxV9ANFDH>
- [25] Dario Kohlbrenner, Christian F. Clarenbach, Adam Ivankay, Lukas Zimmerli, Christoph S. Gross, Manuel Kuhn, and Thomas Brunswiler. 2022. Multisensory Home-Monitoring in Individuals With Stable Chronic Obstructive Pulmonary Disease and Asthma: Usability Study of the CAIR-Desk. *JMIR Human Factors* 9, 1 (2022), e31448. <https://doi.org/10.2196/31448>
- [26] Samantha S.C. Kon, Jane L. Canavan, Sarah E. Jones, Claire M. Nolan, Amy L. Clark, Mandy J. Dickson, Brigitte M. Haselden, Michael I. Polkey, and William D.-C. Man. 2014. Minimum clinically important difference for the COPD Assessment Test: A prospective analysis. *The Lancet Respiratory Medicine* 2, 3 (2014), 195–203. [https://doi.org/10.1016/s2213-2600\(14\)70001-3](https://doi.org/10.1016/s2213-2600(14)70001-3)
- [27] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark Plumbley. 2019. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition (Pretrained Models). <https://doi.org/10.5281/zenodo.3987831>
- [28] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 1 (1992), 159–178. [https://doi.org/10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y)
- [29] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. <http://jmlr.org/papers/v18/16-365.html>
- [30] Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. 2021. LassoNet: A Neural Network with Feature Sparsity. *Journal of Machine Learning Research* 22, 127 (2021), 1–29. <http://jmlr.org/papers/v22/20-848.html>
- [31] Yang Liu, Yu Gu, John Chu Nguyen, Haodan Li, Jiawei Zhang, Yuan Gao, and Yang Huang. 2017. Symptom severity classification with gradient tree boosting. *Journal of Biomedical Informatics* 75 (2017), S105–S111. <https://doi.org/10.1016/j.jbi.2017.05.015>
- [32] Ricardas Marcinkevics and Julia E. Vogt. 2021. Interpretable Models for Granger Causality Using Self-explaining Neural Networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, Virtual Event, Austria. <https://openreview.net/forum?id=DEa4JdMWRHp>
- [33] Ricardas Marcinkevics and Julia E. Vogt. 2023. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery* 13, 3 (2023), e1493. <https://doi.org/10.1002/widm.1493> arXiv:https://www.wiley.com/doi/pdf/10.1002/widm.1493
- [34] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 1 (2008), 53–71. <https://doi.org/10.1111/j.1467-9868.2007.00627.x>
- [35] Madhurananda Pahar, Igor D. S. Miranda, Andreas H. Diacon, and Thomas Niesler. 2021. Deep Neural Network Based Cough Detection Using Bed-Mounted Accelerometer Measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, Toronto, ON, Canada, 8002–8006. <https://doi.org/10.1109/ICASSP39728.2021.9414744>
- [36] Ioannis Pantazopoulos, Kalliopi Magounaki, Ourania Kotsiou, Erasmia Rouka, Fotis Perlikos, Sotirios Kakavas, and Konstantinos Gourgoulis. 2022. Incorporating biomarkers in COPD management: The research keeps going. *Journal of Personalized Medicine* 12, 3 (2022), 379.
- [37] Junfeng Peng, Chuan Chen, Mi Zhou, Xiaohua Xie, Yuqi Zhou, and Ching-Hsing Luo. 2020. A Machine-learning Approach to Forecast Aggravation Risk in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Clinical Indicators. *Scientific Reports* 10, 1 (2020). <https://doi.org/10.1038/s41598-020-60042-1>
- [38] Jean-Louis Pépin, Bruno Degano, Renaud Tamisier, and Damien Viglino. 2022. Remote Monitoring for Prediction and Management of Acute Exacerbations in Chronic Obstructive Pulmonary Disease (AECOPD). *Life* 12, 4 (2022), 499. <https://doi.org/10.3390/life12040499>
- [39] Diego Britto Ribeiro, Aline Carleto Terrazas, and Wellington Pereira Yamaguti. 2022. The Six-Minute Stepper Test Is Valid to Evaluate Functional Capacity in Hospitalized Patients With Exacerbated COPD. *Frontiers in Physiology* 13 (2022). <https://doi.org/10.3389/fphys.2022.853434>
- [40] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Aroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 8 (2017), 913–929. <https://doi.org/10.1111/ecog.02881>
- [41] Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. 2017. Group sparse regularization for deep neural networks. *Neurocomputing* 241 (2017), 81–89. <https://doi.org/10.1016/j.neucom.2017.02.029>
- [42] Robert A. Stockley. 2014. Biomarkers in chronic obstructive pulmonary disease: confusing or useful? *International Journal of Chronic Obstructive Pulmonary Disease* 9, 1 (2014), 163–177.
- [43] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojai, and Emily B. Fox. 2021. Neural Granger Causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 08 (2021), 4267–4279. <https://doi.org/10.1109/tpami.2021.3065601>
- [44] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. <http://www.jstor.org/stable/2346178>
- [45] Stef van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45, 3 (2011), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- [46] Cort J. Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 1 (2005), 79–82. <https://www.jstor.org/stable/24869236>
- [47] Chia-Tung Wu, Guo-Hung Li, Chun-Ta Huang, Yu-Chieh Cheng, Chi-Hsien Chen, Jung-Yien Chien, Ping-Hung Kuo, Lu-Cheng Kuo, and Feipei Lai. 2021. Acute Exacerbation of a Chronic Obstructive Pulmonary Disease Prediction System Using Wearable Device Data, Machine Learning, and Deep Learning: Development and Cohort Study. *JMIR mHealth and uHealth* 9, 5 (2021), e22591. <https://doi.org/10.2196/22591>
- [48] Xiaohan Yan and Jacob Bien. 2017. Hierarchical Sparse Modeling: A Choice of Two Group Lasso Formulations. *Statist. Sci.* 32, 4 (2017), 531–560. <https://doi.org/10.1214/17-sts622>
- [49] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. 2021. Delving into Deep Imbalanced Regression. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual, 11842–11851. <https://proceedings.mlr.press/v139/yang21m.html>
- [50] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (2006), 49–67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x>