Heterogeneity analysis of acute exacerbations of chronic obstructive pulmonary disease and a deep learning framework with weak supervision and privacy protection

Yuto Suzuki, Andrew Hill, Elena Engel, Gregory Lyng, Ann Grancelli, Gabe Lockhart, Farnoush Banaei-Kashani, Russell Bowler

1 Summary

1.1 Background

Chronic obstructive pulmonary disease (COPD) affects 5-10% of the adult US population and is a major cause of mortality. Acute exacerbations of COPD (AECOPDs) are a major driver of COPD morbidity and mortality, but there are no cost-effective methods to identify early AECOPDs when treatment is most likely to reduce the severity and duration of AECOPDs.

1.2 Methods

We conducted the first long-term ($>$ 12 months), real-time monitoring studies of AECOPD with wearable sensors and self-reporting. We applied a deep learning-based autoencoder for feature extractions, then applied K-means clustering to detect heterogeneity. Accordingly, we proposed a weakly supervised active learning framework to develop anomaly detection models for robust identification of early AECOPD, and a clustered federated learning approach to personalize the anomaly detection models for early detection of heterogeneous subtypes of AECOPD. We evaluated this model by comparing it with other unsupervised learning models and federated learning models.

1.3 Findings

We identified two clusters based on the Silhouette score and SHAP analysis. We also found out that a single subject could have exacerbation events from both clusters, indicating that there is not only subject-level heterogeneity but also event-level heterogeneity. Our weakly supervised framework outperformed unsupervised methods by 0.06 in average precision with 25 human annotation labels per subject. Our federated learning framework outperformed standard federated learning methods by 0.14 in F1 score and 0.17 in average precision.

1.4 Interpretation

We showed subject-level and event-level heterogeneity in AECOPD using mobile and wearable device data and developed a practical AECOPD detection framework with limited human annotated labels and keeping data private in each device.

2 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a leading cause of mortality in the United States and acute exacerbations of COPD (AECOPDs) are the major COPD morbidity that leads to excessive mortality, disease progression, and cost of COPD care. In a 5-year longitudinal follow-up study, we found each exacerbation was associated with a 30-75 ml/year decline in lung function, suggesting that AECOPDs contribute to COPD disease progression [11]. Exacerbations also decrease the quality of life [31; 26] and increase mortality [23; 17; 12]. Thus, there are still unmet needs in treating AECOPDs.

Treatment of AECOPDs is typically with corticosteroids and antibiotics. A small study showed that high intensity, supervised monitoring can lead to earlier treatment and better outcomes. Thus, the ability to predict and treat exacerbations early could likely modify the disease progression, reduce cost, and improve quality of life [6; 7]. Other studies have shown that the median time from AECOPD onset to treatment is 4 days and that early treatment is associated with faster recovery and less emergent hospitalization [33]. Unfortunately, 50% to 70% of AECOPDs are unreported [33; 31; 18; 34], suggesting that both late and unreported AECOPDs are unsolved clinical problems. Although it is generally accepted that early treatment of AECOPDs reduces the burden of hospitalization, little is known about the earliest, pre-health care utilization (pre-HCU) characteristics of AECOPDs, and there are no cost-effective methods for early at-home AECOPD detection in the week that precedes a HCU. This gap could be addressed with new wearable sensors and smartphone applications that can generate real-time data streams to monitor COPD patients at home; however, there is a need for automated tools that can process these data and develop personalized risk prediction models for early AECOPD identification. Furthermore, our preliminary work demonstrates that there is considerable heterogeneity in the clinical features that precede AECOPDs before healthcare utilization, which makes it even more difficult to identify AECOPD events before they happen.

Another barrier to improving treatment and outcomes of AECOPDs is difficulty in identifying an early AECOPD while the patient is still at home. The major drawback to home-based AECOPD disease management programs are that they require active participation and expensive personnel costs. An alternative solution to these programs could be automated data collection with wearable sensors and smartphone apps coupled with artificial intelligence/machine learning (AI/ML) based monitoring. However, the potential of clinical tools utilizing state-of-the-art AI/ML techniques has not yet been explored. In this study, We aimed to analyze the heterogeneity of AECOPDs and evaluate a deep learning

framework for anomaly detection with minimum labeled data and privacy data protection.

The goal of this project is to identify the heterogeneity of COPD in data coming from wearable devices and introduce methods for advanced and robust detection of early clinical features in heterogeneous subtypes of AECOPD. If successful, our proposed methods can also be adopted for the early detection of acute exacerbations in other chronic diseases.

3 Methods

3.1 Recruitment of subjects

Subjects were recruited from a single special clinic and through online advertising. Eligibility included a physician diagnosis of COPD and an AECOPD treated with corticosteroids or antibiotics in the preceding 12 months in order to enrich for future exacerbations. This study was approved by an institutional review board and all subjects gave informed written consent.

3.2 Study procedures

193 subjects were screened and 73 subjects consented. After completing informed consent, 50 subjects were enrolled in a minimum 1 week run in period in which they were asked to complete the The Exacerbations of Chronic Pulmonary Disease Tool (EXACT) on their smartphone (Apple or Android). The EXACT is a 14-question patient-reported outcome (PRO) measure for standardizing the symptomatic evaluation of AECOPDs and acute exacerbations of chronic bronchitis (AECB) in natural history studies and clinical trials [20]. Fifty subjects completed the EXACT for at least 6 days in the first week of the run-in and were mailed the following sensors: a smart watch (Fitbit Sense 2 [1] or Apple Watch series 6 [2]), inhaler sensors (Propeller Health [3]), and a smart thermometer (Kinsa Health [4]). Subjects were then guided through the installation of the devices and applications over the phone and asked to wear the smartwatch and answer the smartphone survey daily.

3.3 Event adjudication

Subjects were encouraged to self-report AECOPDs, but were also called every 4 months by a research coordinator to complete an exacerbation questionnaire. If necessary, AECOPD medical records were obtained, redacted, and reviewed by a pulmonologist, who adjudicated (1) whether the event was an AECOPD and what date it started; and (2) the severity of the event: severe being defined as requiring hospitalization or emergency room; moderate being not severe, but requiring HCU with corticosteroids or antibiotics; or mild, not requiring HCU or prescribed medications.

3.4 Data Collection and Processing

Subjects were asked to complete the digital EXACT at a consistent self-selected time during the evening. For subjects with an iPhone, data were collected using a bespoke iOS app (www.Circitores.com). For other subjects, we collected smart watch data using Fitabase and the EXACT using RedCAp. Propeller data were collected from the Propeller research portal. Repeat actuation occurring $<$ 2 minutes apart were considered as a single inhaler use event [9]. Physiologic data including heart rate (HR), steps, EXACT, inhaler usage, and adjudicated AECOPD events were then temporally aligned.

3.5 Predictor Variables

The predictors/features we used are summarized in Table 1, 2, and 3. We converted all features into a daily basis by averaging for the purpose of computational efficiency.

Table 1. Description of data features

Feature	Description
EXACT	Development of the EXAcerbations of Chronic Obstructive Pulmonary Disease Tool
Oxygen Saturation	A measure of how much hemoglobin is currently bound to oxygen compared to how much hemoglobin remains unbound.
Rescue Inhaler Use	Medication for relieving symptoms
Controller Inhaler Use	Medication for prevention
Heart Rate	The number of times your heart beats per minute
Steps	How many steps you take each minute
Calories	Total calories burned for the min [20]

Table 2. Details of data features

Feature	Domain	Source	Type of Data(Range)	Temporal Resolution
EXACT	Symptoms	App Survey	Score(0-100)	Daily
Oxygen Saturation	Physiology	App Survey	SpO2(0-100%)	Daily
Rescue Inhaler Use	Medication	Propeller	Puffs per day	Daily
Controller Inhaler Use	Medication	Propeller	Puffs per day	Daily
Heart Rate	Physiology	Fitbit or Apple watch	Average beats/min	min
Steps	Activity	Fitbit or Apple watch	steps	min
Calories	Activity	Fitbit or Apple watch	calories	min

3.6 Data analysis

Our analysis process included data preprocessing, dimensionality reduction, clustering, and SHAP values analysis [22] as described below.

Table 3. Data statistics

Feature	mean ± standard deviation	Missing rate
EXACT	38.5±12.7	23.7
Oxygen Saturation	93.0±2.80	52.4
Rescue Inhaler Use	1.56±3.40	0
Controller Inhaler Use	1.67±2.73	0
Heart Rate	76.4 ±13.3	54.7
Steps	1.99±8.62	46.4
Calories	1.20±0.85	46.4

Missingness and training dataset. We selected 8 patients who had at least 2 exacerbation events, in order to make both train datasets and test datasets for anomaly detection experiments. Thereafter, we converted all features into daily basis data by averaging for computational efficiency. We dropped consecutive missing values with > 1 consecutive day of missing data for heart rate and steps while we dropped missing values with > 3 days for EXACT and oxygen saturation, and we imputed missing values with linear interpolation otherwise. We standardized each feature after imputation.

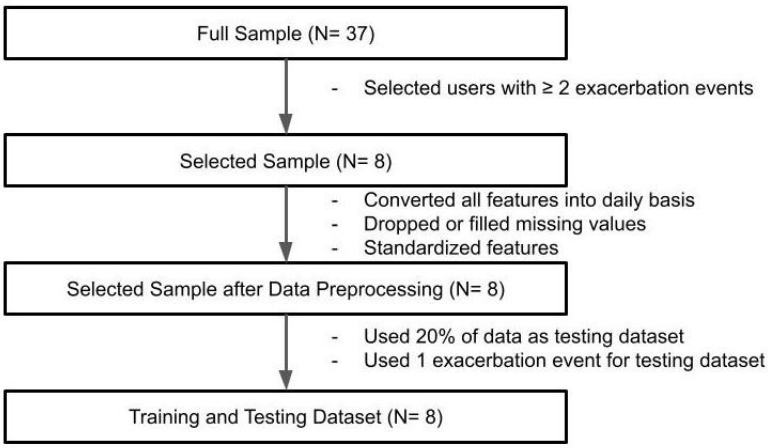

Fig. 1. Data preprocessing flow

Dimensionality reduction. Because daily-based data is still high dimensional and it is hard to compute the distance between those data, we reduced data dimensionality by using an LSTM-based autoencoder. We converted all features into 8-dimensional vectors. (Figure 2) When we focused on patient-level heterogeneity in this section, we selected one consecutive pre-exacerbation event data group from each patient to observe data trends related to exacerbation events.

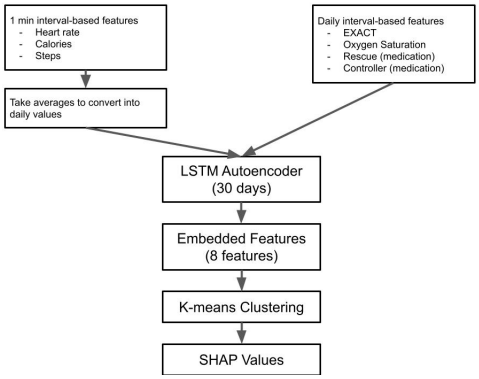

Fig. 2. Dimensionality reduction flow

Clustering. We applied K-Means clustering [21] to the LSTM [15] embedded data and analyzed silhouette score [29] and analyzed each cluster by game theory-based Shapley Values (SHAP Values) [22]. We selected the number of clusters $k = 2$ because 2 clusters led to the highest silhouette score (5).

Early AECOPD prediction. We trained an LSTM-based classification model to test whether it could detect anomalies (AECOPD) from in the 0-30 days prior to HCU from an AECOPD. We designed an anomaly detection framework for the purpose of early detection of AECOPD. The overview of our framework is summarized in Figure 3. This framework consists of two components. The first component is an active learning-based anomaly detection component in each patient. This allows each subject to train an anomaly detection model within their device without any privacy concerns or communication costs with a central server. The second component is Federated Learning (FL) [24]. FL allows each patient to communicate with a central server and share knowledge with other patients without sharing raw data, thus keeping individual data private. This is possible by sharing only trained model parameters. Because we had observed AECOPD heterogeneity by subject , we considered heterogeneity during the server aggregation. We applied the same data preprocessing process as figure 1. First, we randomly selected one exacerbation event per patient for test data, and then we assigned 80 % of the data as the training dataset and 20 % as a test dataset. We used a modified version of Label-Efficient Interactive Time-Series Anomaly Detection (LEIAD) [13]. This framework utilizes active learning to maximize the model performance with a limited number of human-annotated label data with an ensemble of unsupervised anomaly detection, which increases the robustness of the active learning label generation process. Our model considers multi-variate time series data and uses an LSTM-based deep learning model to be able to capture complex multi-variate data structures. In our experiment, we added 5 true labels after each iteration.

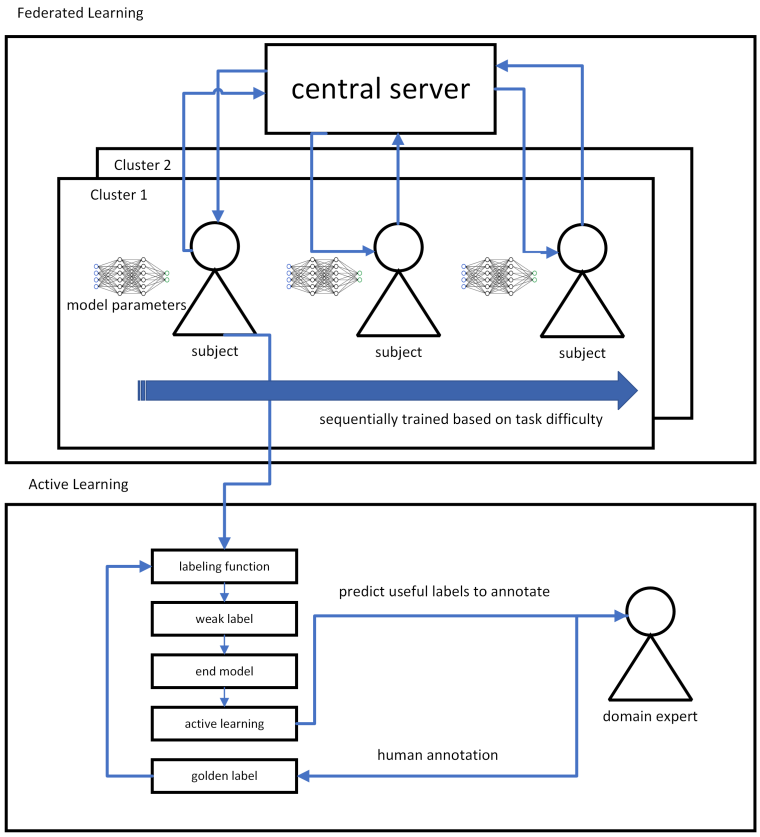

Fig. 3. Framework Overview

Federated Learning (FL). FL is a relatively new machine learning paradigm where multiple machine learning models train collaboratively without sharing raw data. This is possible by sharing only parameters with a central server, and the server aggregates those parameters by a computational function, such as averaging or weight averaging, to share knowledge implicitly. One of the main challenges in applying FL to our setting is data heterogeneity. As we have seen, each patient, and even each exacerbation event, could be heterogeneous. Hence, aggregating all patients' models into a single global model could be detrimental. The other potential issue is paucity of data. Each patient has only 2-4 exacerbation events, which could lead to poor performance in each model, and aggregating poor local models does not usually work.

To overcome these challenges, we proposed a novel deep learning-based method with three components. The first one is the daisy-chain algorithm [16] which allows a central server to send a global model to one subject, train the model,

send it back to the server, and iterate this process for another subject. This replicates training using all raw data, which works well when data are limited. The second component is clustering with a novel distance measure such as cosine similarity of model parameters as a distance measure for clustering. However, in our case, since the data from each patient device is limited, each local model might be poor, which might not represent data characteristics well. Therefore, we used model loss as a distance measure. First, we randomly select one subject and train a local model. Thereafter, we send the model to all other patients and compute model loss, and if the model loss is smaller than a threshold, we make a cluster for them. If a model loss is smaller, that means that a model that works for a local dataset could work for another dataset, implying similar data distribution. In addition to that, this can measure data characteristics more directly than using cosine similarity for model parameters. That is why our similarity measure is more suitable for limited datasets than cosine similarity would be. The third component is curriculum learning [32]. Curriculum learning is a machine learning technique where we first train a model using easier data/tasks and later train a model with more difficult data/tasks, which is similar to how human brains work. In order to optimize the order in the daisy-chain algorithm, we used curriculum learning.

4 Results

We selected the number of clusters $k = 2$ for patient-level clustering because it achieved the highest silhouette score and steep slope in the Elbow method (Figure 4 and Figure 5). The silhouette score near 0.5 shows that the 2 clusters are somewhat separate, which may indicate that there is some heterogeneity among data trends before exacerbation events among patients.

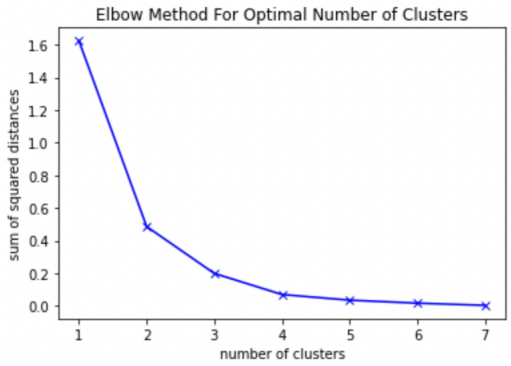

Fig. 4. Elbow method for patient-level clustering

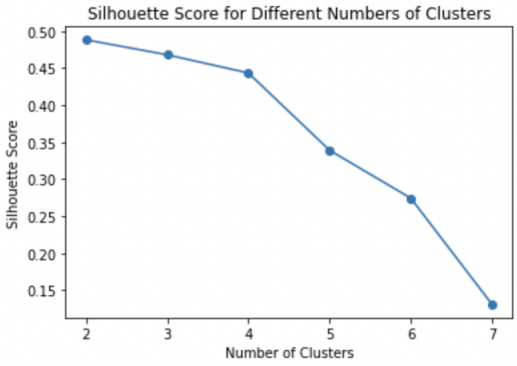

Fig. 5. Silouette score for patient-level clustering

In addition to the silhouette score, we analyzed each cluster by game theory-based Shapley Values (SHAP Values) to observe the characteristics of each cluster. High SHAP values of features mean that those features are more informative as predictive variables. In our subject-level analysis, we visualized SHAP values of clustering. Hence, each SHAP value shows the contributions of each feature to clustering. Figure 6 shows absolute values of SHAP values for 30 days before an exacerbation event. Figure 6 implies that predictive variable data has heterogeneity from day 16 to day 30 where day 31 is the day when an exacerbation event occurs. SHAP values increase over time, meaning that data closer to an exacerbation event has more impact on clustering. This means that data trends could show more unique trends in each cluster as it gets closer to an exacerbation event. This shows the heterogeneity of exacerbation events and the necessity of anomaly detection considering the heterogeneity among patients.

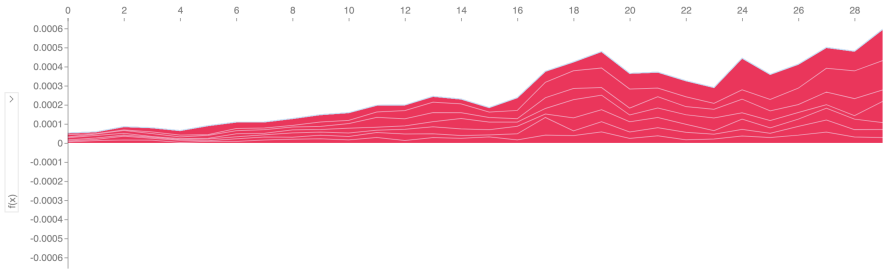

Fig. 6. Shap values for 30 days before an exacerbation event (subject-level)

Figure 7 shows how each feature contributes to clustering. Figure 7 shows that one cluster (namely, cluster 0) has low steps, high EXACT, high heart rate,

high control medication, high oxygen saturation and the other cluster (namely cluster 1) has low exact, high oxygen, and low rescue. These results with the silhouette score analysis indicate that data trend before an exacerbation event has heterogeneity and the heterogeneity increases over time as it gets closer to an exacerbation event.

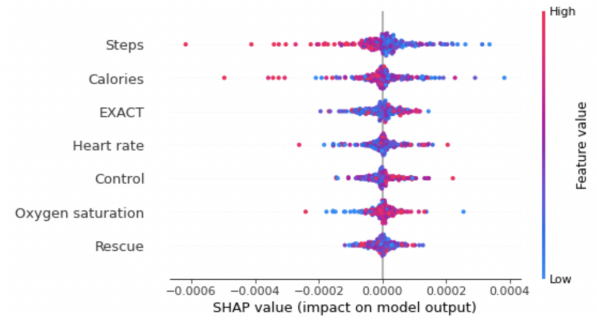

Fig. 7. Shap values (subject-level)

Our SHAP analysis for clustering exacerbation events regardless of subjects (event-level) is in figure 8 and 9. Figure 8 shows a similar trend as subject-level heterogeneity analysis, which shows an increase in heterogeneity over time. Figure 9 shows that one cluster (cluster 0) has low heart rate, high control medication, high steps, high EXACT, high rescue medication, high calories and low oxygen saturation, and the other cluster (cluster 1) has opposite characteristics.

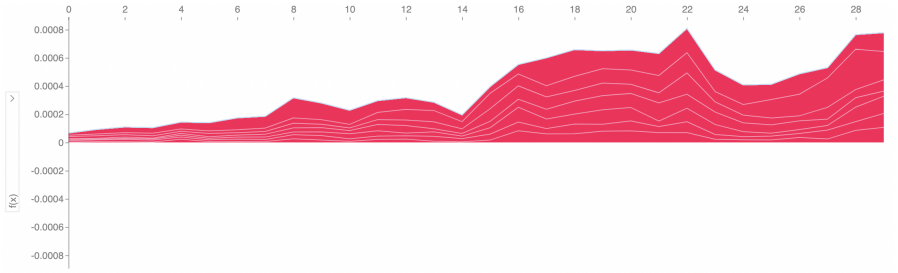

Fig. 8. Shap values for 30 days before an exacerbation event (event-level)

Table 4 shows which cluster each exacerbation event belongs to. 5 out of 8 patients have 2 clusters in a single patient, indicating that there could be heterogeneity even among a patient.

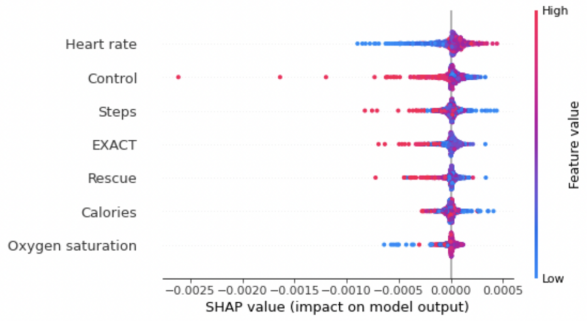

Fig. 9. Shap values (event-level)

Table 4. Characteristics of study participants

Patient ID	cluster id (event 0)	cluster id (event 1)	cluster id (event 2)	cluster id (event 3)	cluster id (event 4)
0	1	1			
1	0	0			
2	1	1			
3	0	1			
4	1	0	1		
5	1	0	1		
6	0	1	0	1	0
7	0	1	0		

Figure 10 and figure 11 are data samples. In figure 10, we can observe similar data trends before exacerbation events (yellow areas) where EXACT, rescue medications, heart rate, and calories increase. This implies that this subject does not have heterogeneity in AECOPD. In figure 11, we can observe high oxygen saturation, high calories before one exacerbation event whereas we can observe low oxygen saturation, low heart rate, low steps and low calories before the other exacerbation event, indicating there is heterogeneity in AECOPD in a single subject. This example also shows the existence of event-level heterogeneity.

Figure 12 and 13 show performance comparison in F1 and Average precision regarding LEIAD and other unsupervised learning models in the test dataset. They show average results among 8 patients. We observed that our modified LEIAD achieved the highest performance in Average Precision after 3-5 iterations. This indicates that LEIAD could outperform unsupervised models only with 15-25 human-annotated labels per patient. This was possible because LEIAD identifies which label data we need to improve model performance with a limited number of labeled data. our modified LEIAD achieved the second-best result in F1 score and we noticed the performance drop after 25 human-annotated labels. This could be explained by that there is heterogeneity in labels and adding those true heterogeneous labels could be noise.

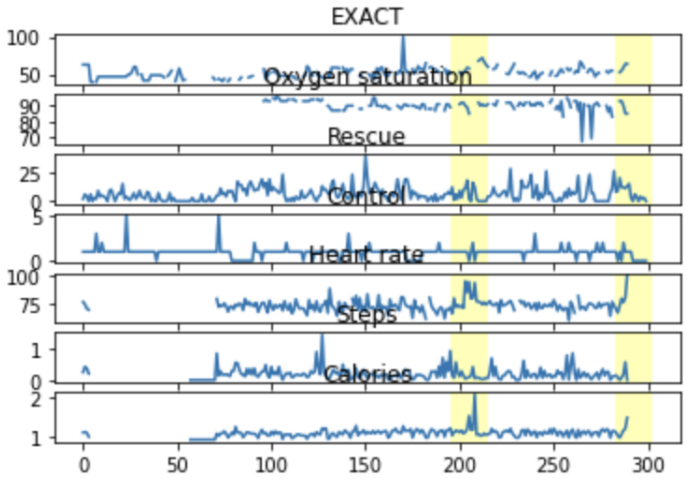

Fig. 10. Sample data 1 (no heterogeneity in AECOPD)

Table 5 shows the results of various FL algorithms for our dataset. We first used LEAID with 25 human annotations and then applied each FL algorithm to test practical applications. LEIAD + Standalone, LEIAD + FedAvg, and LEIAD + FedDC achieved the same results, which indicates that FL does not improve the results. This could be explained by that aggregating all local models is not beneficial due to data heterogeneity. When we add clustering and curriculum learning, those models increase model performance, indicating that aggregating models among clusters and curriculum learning would be effective when there is heterogeneity and data is limited. This is because sharing common knowledge only within the same clusters is beneficial in patients in the same clusters. The biggest performance gain was possible by clustering, also showing heterogeneity in AECOPD.

Table 5. Model performance comparison

Model	F1	Average Precision
LEIAD+Standalone	0.22 ±0.00	0.33±0.03
LEIAD+FedAvg	0.22 ±0.00	0.36±0.03
LEIAD+FedDC	0.23 ±0.00	0.35±0.01
LEIAD+FedDC+ Clustering	**0.36 ±0.04**	0.50±0.07
LEIAD+FedDC+ Clustering + Curriculum Learning	**0.36 ±0.02**	**0.53±0.03**

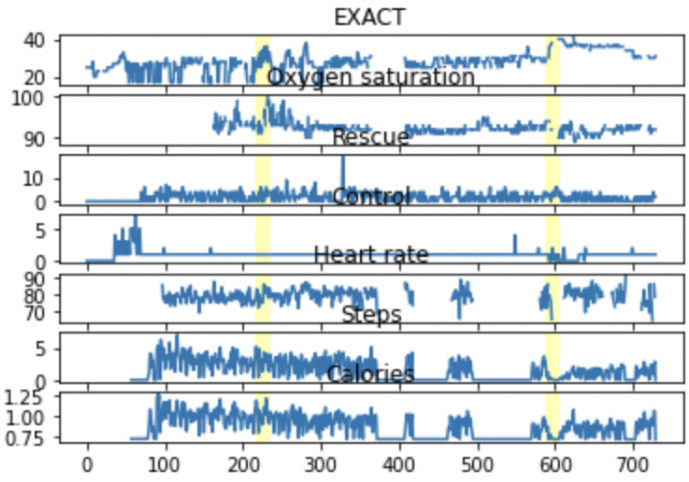

Fig. 11. Sample data 2 (heterogeneity in AECOPD)

5 Discussion

In this work, we analyzed heterogeneity in AECOPD at the subject-level and event-level, and we proposed a practical anomaly detection framework based on weakly supervised learning and federated learning. We found that the data trend right before an exacerbation event has 2 clusters, which indicates that there is AECOPD heterogeneity. We also observed that some subjects have exacerbation events from both clusters, meaning that there is heterogeneity even among a single subject. In order to address this heterogeneity and lack of data, we proposed an anomaly detection framework utilizing weakly supervised learning and federated learning. Our weakly supervised learning method helps medical experts to make annotated data efficiently so that we are able to maximize model performance with limited human annotated labels. Our Federated Learning method allows us to make patient-level clusters and share knowledge within clusters without sharing raw data (for privacy protection). Our experiments show the potential of the practical application of the framework in a real-world setting where medical experts add a small number of annotated labels, train models in each patient device, and apply federated learning in a central server for further improvement considering heterogeneity.

We explored heterogeneity in AECOPD and found important insights about patient-level and event-level heterogeneity. We found two clusters. Interestingly enough, a single subject could have exacerbation events from both clusters, meaning that this AECOPD heterogeneity is not necessarily subject-specific, which makes early detection more challenging. SHAP values also showed that the heterogeneity increased over time, starting to show a signal 10-20 days before an exacerbation event and becoming stronger over time. This may show the

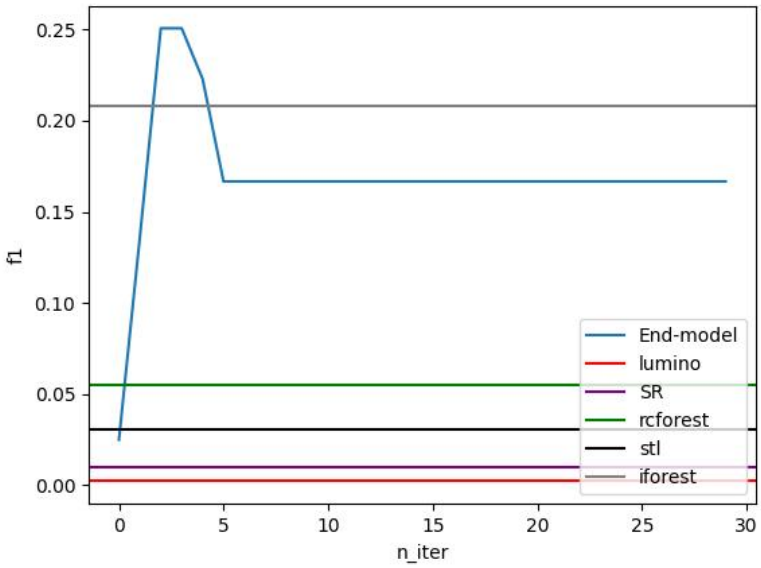

Fig. 12. Average F1 score among patients of LEIAD and other unsupervised methods

possibility of earlier detection, however, more exacerbation training data sets will be needed for earlier detection.

We proposed a practical anomaly detection framework and showed potential through several experiments. Each subject utilizes active learning where an ensemble of unsupervised models suggests the most informative data to annotate. We found that our model outperformed baseline solutions based on unsupervised anomaly detection only with 15-25 annotated events, which shows practical use with high efficacy. We modified the existing framework to adjust to our multivariate and limited data scenarios by including an over-sampling process, deep learning-based feature extraction, and an ensemble of all features. We also used federated learning in order to share knowledge among patients without sharing raw data. FL achieves this goal by only sharing model parameters. In addition to that, we applied clustering based on model loss where we trained one model with one patient data and computed model loss for each patient and group patients with small loss. Thereafter, we used a daisy-chain algorithm and curriculum learning to address the problem of a limited number of events. Training a model sequentially, especially in order of task difficulty, replicates a situation where we train a single model with all local data, which leads to model performance improvement in a limited number of data.

In addition to early detection, efficient home monitoring could ameliorate disparities in health care. There are also health disparities in COPD (and in

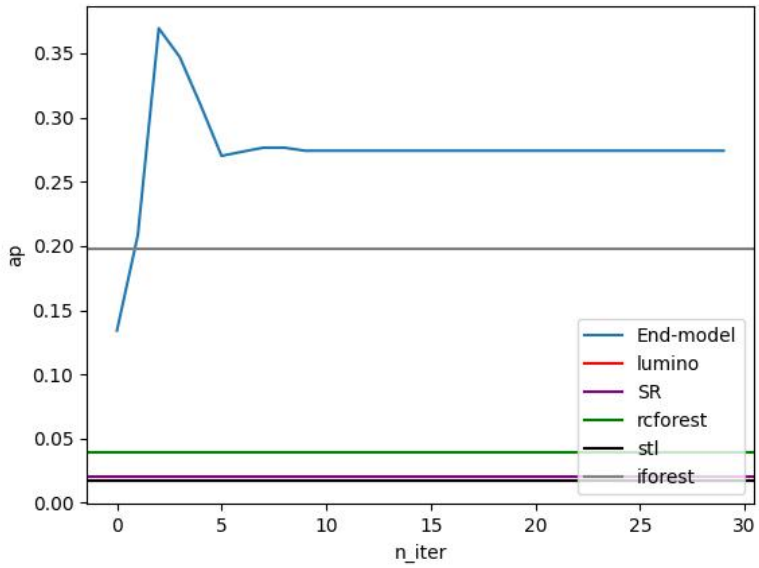

Fig. 13. Average Average precision among patients of LEIAD and other unsupervised methods

turn, in AECOPD). The three leading causes of death in rural areas (heart disease, cancer, and stroke) have been decreasing, yet the rural pre-COVID mortality rate for COPD has been increasing [25]. Rural patients have significantly higher COPD prevalence, Medicare hospitalization for COPD, and COPD related deaths [28; 27; 10]. Black Americans are also a group for which there is poor access to COPD special care and worse quality of life with exacerbations [14] and also reduced access to special care for chronic lung diseases compared to white Americans [8].

There are some limitations in this work. First, we converted minute-based data coming from Fitbit and Apple Watch, such as heart rate and steps, however, this could have led to some information loss, leading to model performance degradation. On the other hand, training a model using minute-based data requires much more computational power, so our future work is to develop a computationally efficient method so that we can realize it in each patient's device. Second, the variety of data features is still limited. For instance, we are considering collecting temperature and air quality data using GPS. Third, since most of the patients have only 2 exacerbation events, they have only 1 exacerbation event for training data, hence, we were not able to develop a framework and test event-level heterogeneity. Our current framework only considers patient-

level heterogeneity, therefore, our future work is to collect more data per patient and extend our method to handle event-level heterogeneity.

Our work detected heterogeneity at the subject-level and event-level and proposed a practical AECOPD detection framework with limited human annotations and privacy protection. We identified two distinct clusters and found that a single patient could have an exacerbation event from both clusters. Our AECOPD detection framework incorporates weakly supervised learning and federated learning with novelty in accommodating multi-variate time series data, clustering based on local data implicitly, and applying FL in the order of task difficulty. We showed improvement from our baseline solutions and showed potential for practical medical applications. Our research shed light on new research directions in AECOPD heterogeneity and how to address it in real-world applications with recent artificial intelligence technologies.

6 Supplemental Document

7 Materials and methods

7.1 Data Collection

Respiratory symptoms such as dyspnea, wheezing and sputum production define AECOPDs, but are traditionally ascertained retrospectively. In our pilot studies we used diaries such as the EXAcerbations of Chronic Pulmonary Disease Tool (EXACT), which more reliably assesses frequency, severity, and duration of COPD exacerbations [20] and can be administered electronically using smartphones [9; 30]. Physical activity and rescue medication use are also two important factors that are strong predictors of COPD outcomes but are less reliable when self-reported. Recent technological advances now allow real time passive electronic collection of activity, heart rate, pulse oximetry, temperature, and inhaler use (see Figure 1). These sensors are particularly attractive to use because they passively collect and transmit data through smartphone. Using aforementioned data collection means, we have conducted two pilot studies of 3-week and 12+ month enrollment duration to collect data and evaluate efficacy of basic methods as well as challenges in early detection of AECOPD. First, in our recently published 3-week pilot study , 184 COPDGene subjects were enrolled over six months at one clinical center. The subjects used an Android smartphone to complete an eDiary (EXACT), continuously wore an activity monitor (ActiGraph [5]) for monitoring steps and calories, and used real-time rescue inhaler monitoring (Adherium SmartInhaler) as well. Although the three-week pilot was not powered to detect health care utilization exacerbations, we detected 9 EXACT defined exacerbation events. We also identified novel rescue inhaler patterns that were independent of total inhaler doses per day and found that the EXACT score was positively associated with COPD progression and daily rescue inhaler use [9]. These publications demonstrated that some exacerbation events have prodromal symptoms that can be detected with real-time monitors while

the subject is at home, but also suggest the need for the development of machine learning algorithms for better real time detection of events with integrated data. Full compliance (all three devices) over three weeks was 98% (180/184). Post study interviews with the subjects indicated that they did not find the study burdensome because most of the data are collected passively, other than the eDiary (which is short, and our subjects do not report as excessively burdensome. However, academic reviewers were skeptical that subjects would continue to wear devices and answer eDiary over 12 months. Therefore, for a second pilot study we enrolled new subjects from the community for a 12-month study. To date, this pilot study has enrolled 30 subjects at a single site (National Jewish Health) and only one subject has withdrawn from the study. Nearly all subjects have chosen to stay in the study for more than 12 months. In the 12-month pilot we used Fitbits instead of Actigraphs and we used Propeller instead of Adherium inhaler sensors. The COVID pandemic spurred additional innovations in the 12-month pilot including: electronic consent, remote enrollment, and addition of temperature and oxygen saturation to the sensor streams. In 2021 we added COVID surveillance in our three month follow up; however, none of the AECOPDs were due to COVID, which was consistent with marked reduction of AECOPDs in general because COPD patients were self-isolating in 2020 and 2021. In January 2022 Apple began providing us with the latest Apple watches (which can collect oxygen saturation data) and in April 2022 Apple approved our iPhone app for real-time collection and streaming of eDiary, inhaler usage and sensor data (steps, heart rate, oxygen saturation, and temperature) through Apply HealthKit to a protected cloud bucket (see Figure 2). Compliance with both eDiary and devices has been excellent in the 12-month pilot. For instance, 83% of subjects answered the EXACT survey on more than 75% of days and 74% of subjects wore the wrist device (Fitbit) more than 85% of days.

To begin to address this problem, we have conducted preliminary home-based real-time monitoring studies of high-risk COPD patients using wearable sensors. In our pilot work, we demonstrate a real-time, home-based monitoring approach using eDiaries such as the EXAcerbation of Chronic Pulmonary Disease Tool (EXACT) [19], real-time monitoring of inhaler and wrist sensors [9] and residential bioaerosol sampling, is feasible (>80% compliance) and can be cost-effective over both short term (3-weeks) and long-term (>12 months). These preliminary studies are the first to include and integrate real-time reporting of activity, heart rate, SpO2, temperature, inhaler use, and symptoms, and they have allowed us to develop and test basic unsupervised methods for early detection of AECOPD.

7.2 Imputation

One issue with the collected sensor data was considerable missing data. Participants may occasionally forget to utilize one of the devices or answer the survey, resulting in missing values for one or more sensor streams.

Missing Values Table 14 shows the pattern of missingness in the dataset. We observe heterogeneity of patterns of missingness for features, patients, and

timing. Table 15 shows the correlation of missingness between every 2 features. It shows high correlation between EXACT and oxygen saturation, rescue and control medications, and heart rate and steps. All these pairs are collected from the same methods/devices, therefore, we could see that the way of collecting data has a big impact on missingness.

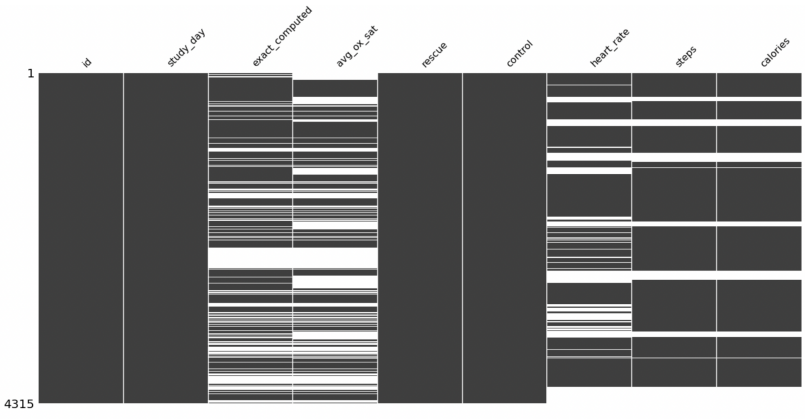

Fig. 14. Missing data distributions

Dropping Missing Values We used 2 strategies to handle missing values. The first strategy is to drop missing values based on a threshold. We need to find a threshold which does not drop too many missing values, but also removes data with too many missing values. In order to find an appropriate threshold to meet this requirement, we visualized the number of missing values for each consecutive day in figure 16 and 17 for oxygen saturation feature. In this case, we selected 4 days as a threshold and dropped missing values with more than 4 consecutive days missing values. This way, we can still keep enough data, but still remove data with too many missing values. In the same way, we selected 4 days for EXACT, 1 day for heart rate, and 1 day for steps as a threshold. Regarding rescue and control medications, we did not remove any missing values and replaced them with 0 because we assumed that missing records in medications were most like that there was no medication on that day.

Imputation Method The other method to handle missing values is imputation. We evaluated 4 different imputation methods. The first one is linear imputation. The second one is deep learning-based forecasting in each patient's data. We train LSTM-based forecasting models, which take 10 days data on all features and make predictions for all features of the next day. The third

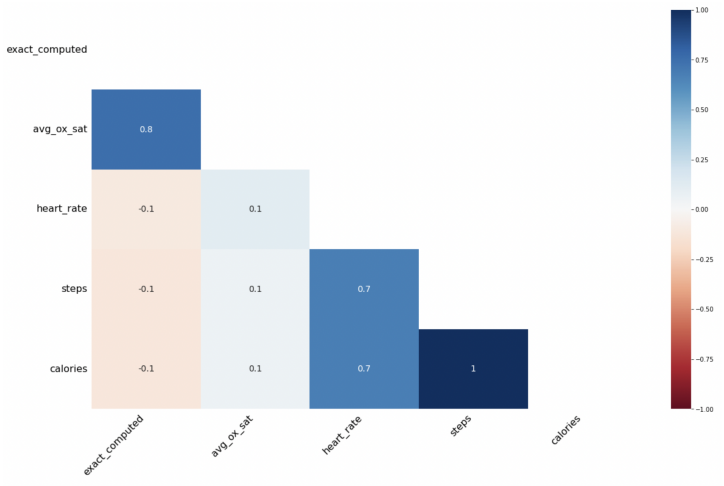

Fig. 15. Heatmap of missing values

one is based on the same LSTM forecasting model, but it uses FedAvg algorithm among patients. The last one is also based on the same LSTM forecasting model, but it uses our FL algorithm (Daisy-chain algorithm with clustering and curriculum learning) among patients. Figure 18 to 22 shows sample data after each imputation method is applied, and table 6 shows the comparison of the model performance. We found all results produced comparable results and did not see significant differences. This is mostly because we already dropped long consecutive missing values. Therefore, we used linear imputation for the rest of the experiments for computational efficiency.

Table 6. Imputation method comparison

Imputation Method	F1	Average Precision
Linear interpolation	0.14 ±0.04	0.23±0.06
LSTM forecasting (Local training)	0.08 ±0.04	0.19±0.06
LSTM forecasting (FedAvg)	0.14 ±0.11	0.22±0.09
LSTM forecasting (Our model)	0.12 ±0.03	0.19±0.05

7.3 Early detection

It is also worth noting that this heterogeneity shows the potential of early detection of AECOPD because we might be able to catch a warning sign of anomalies starting from 30 days before an exacerbation event regarding survey and medication data, and 17 days before an exacerbation event regarding Fitbit data.

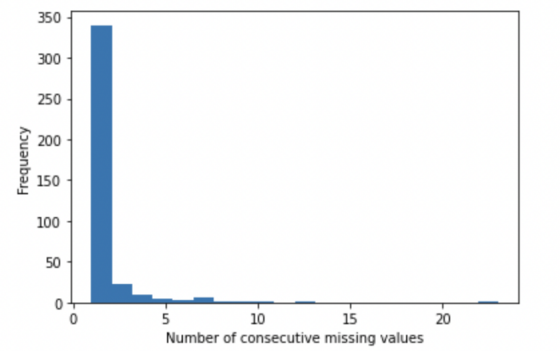

Fig. 16. Consecutive missing values of oxygen saturation

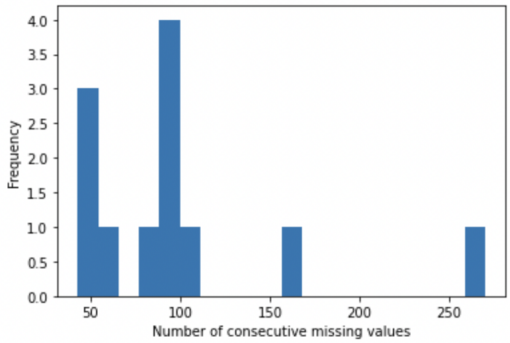

Fig. 17. Consecutive missing values of oxygen saturation

Therefore, we conducted an experiment to see whether a deep learning model can detect AECOPD earlier. Figure 23 shows F1 score based on the test dataset for AECODP prediction 0-30 days prior to an exacerbation event. Figure 23 indicates that there is no consistent trend except that a model achieved the highest F1 score when it predicts from one day before AECOPD. This could be explained by that we might need more data to detect anomalies from earlier days since the sign of anomalies is more subtle.

7.4 Anomaly Detection Framework

Weakly Supervised Learning We performed a modified version of Label-Efficient Interactive Time-Series Anomaly Detection (LEIAD) [13]. LEIAD first creates a label function to predict labels for unlabelled data using an ensemble of unsupervised models. In addition to this, we added an ensemble process of

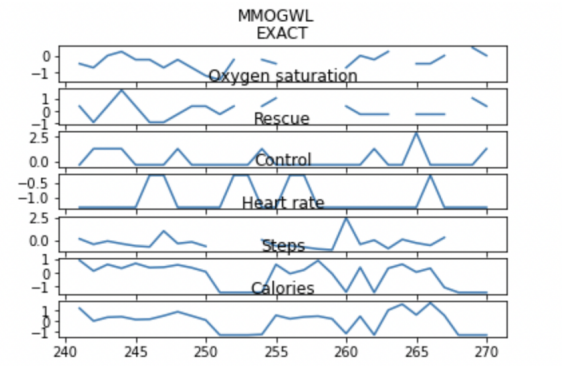

Fig. 18. Sample of Data (without imputation)

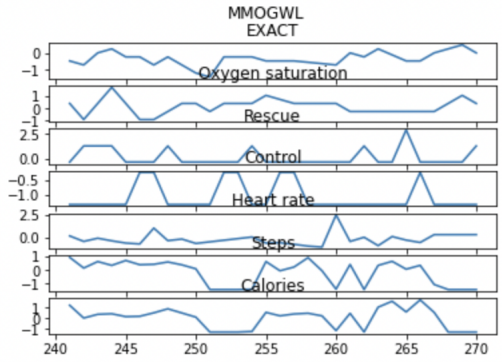

Fig. 19. Sample of Data (with linear interpolation)

multi-variate time series features. We made anomaly labels in cases when unsupervised methods based on more than one feature detect anomalies. Thereafter, we iterated the process of active learning where domain experts annotated labels, updated label functions, and trained the model. We set 5 label annotations as a minimum unit and tested the impact. Another improvement is that we use deep learning-based feature extractors, which is LSTM, instead of feature engineering such as moving average in LEIAD. This is because LSTM can do feature extractions through a learning process and this is especially useful when data has high dimensional (in our case, we have multi-variate time series data, which is high dimensional).

Federated Learning We further applied Federated Learning (FL) to share common knowledge among patients without sharing raw data. FL is a relatively new machine learning paradigm where multiple clients (e.g., smartphones or

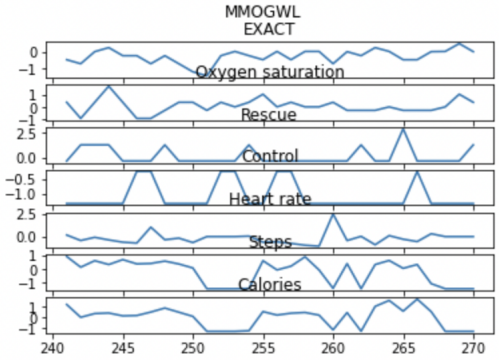

Fig. 20. Sample of Data (with local LSTM)

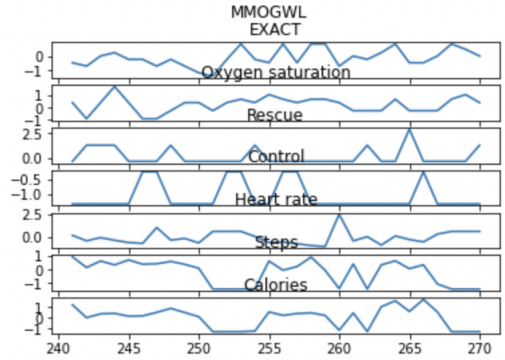

Fig. 21. Sample of Data (with fedavg LSTM)

hospitals) train machine learning models collaboratively while keeping their raw data private. The most common FL approach is FedAvg where each local client sends model parameters to a central server and the server sends back the average model parameters to each local client. Applying FedAvg is problematic for our data because of heterogeneity and limited data. With data heterogeneity among each user, the optimal model for each local user is different and a single global model aggregating all local models could be far from each local optimal model. In addition, since each local user lacks data in our scenario, each local user could train poor models. Aggregating local poor models is known not to work well. In order to address these challenges, we developed a novel FL approach where we combine FedDC, clustering, and curriculum learning. FedDC is our baseline model, which is recent work for small datasets. FedDC uses daisy-chain, where a central server receives one local model, and sends it to another client, the local client trains the model and sends it back to the server. This way, a

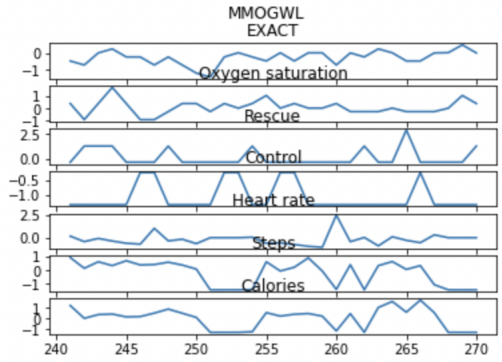

Fig. 22. Sample of Data (with feddc LSTM)

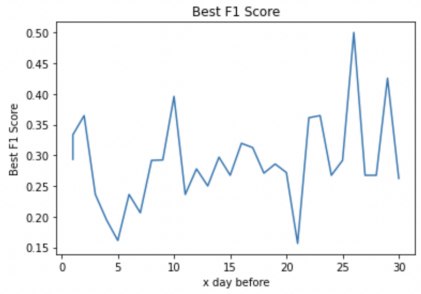

Fig. 23. Best F1 Score based on various timing of early detection

model trains with all local data sequentially. FedDC could still suffer from data heterogeneity, therefore, we apply clustering first. Instead of using a common distance measure of model parameters such as cosine similarity, we used model loss in order to measure distance based on data more directly. We first randomly select one client, train a model, send it to all other clients, compute model loss based on the locally trained model, and make a cluster if a model loss is smaller than a threshold. We iterate this process until all clients belong to some clusters. In addition to this improvement, we hypothesized that an order of daisy-chain could be crucial, so we applied curriculum learning, which is a machine learning technique where a model is trained with easy data/task first and then gradually with more difficult data/task in the same way as a human learns to master a new skill. We ordered clients within a cluster based on a model loss so that we could train from easier clients.

Bibliography

[1] https://www.fitbit.com/global/us/products/smartwatches/sense2

[2] https://support.apple.com/kb/SP826?locale=en$_US$

[3] https://propellerhealth.com/

[4] https://www.kinsahealth.com/

[5] https://theactigraph.com/

[6] Chronic obstructive pulmonary disease: management of adults with chronic obstructive pulmonary disease in primary and secondary care. National Clinical Guideline Centre (2010)

[7] Global initiative for chronic obstructive lung disease (2011)

[8] Bellinger, J.D., Hassan, R.M., Rivers, P.A., Cheng, Q., Williams, E., Glover, S.H.: Specialty care use in us patients with chronic diseases. International Journal of Environmental Research and Public Health **7**(3), 975–990 (2010)

[9] Bowler, R., Allinder, M., Jacobson, S., Miller, A., Miller, B., Tal-Singer, R., Locantore, N.: Real-world use of rescue inhaler sensors, electronic symptom questionnaires and physical activity monitors in copd. BMJ Open Respiratory Research **6**(1), e000350 (2019)

[10] Croft, J.B., Wheaton, A.G., Liu, Y., Xu, F., Lu, H., Matthews, K.A., Cunningham, T.J., Wang, Y., Holt, J.B.: Urban-rural county and state differences in chronic obstructive pulmonary disease—united states, 2015. Morbidity and Mortality Weekly Report **67**(7), 205 (2018)

[11] Dransfield, M.T., Kunisaki, K.M., Strand, M.J., Anzueto, A., Bhatt, S.P., Bowler, R.P., Criner, G.J., Curtis, J.L., Hanania, N.A., Nath, H., et al.: Acute exacerbations and lung function loss in smokers with and without chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine **195**(3), 324–330 (2017)

[12] Groenewegen, K.H., Schols, A.M., Wouters, E.F.: Mortality and mortality-related factors after hospitalization for acute exacerbation of copd. Chest **124**(2), 459–467 (2003)

[13] Guo, H., Wang, Y., Zhang, J., Lin, Z., Tong, Y., Yang, L., Xiong, L., Huang, C.: Label-efficient interactive time-series anomaly detection. arXiv preprint arXiv:2212.14621 (2022)

[14] Han, M.K., Curran-Everett, D., Dransfield, M.T., Criner, G.J., Zhang, L., Murphy, J.R., Hansel, N.N., DeMeo, D.L., Hanania, N.A., Regan, E.A., et al.: Racial differences in quality of life in patients with copd. Chest **140**(5), 1169–1176 (2011)

[15] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[16] Kamp, M., Fischer, J., Vreeken, J.: Federated learning from small datasets. arXiv preprint arXiv:2110.03469 (2021)

[17] Kim, S., Clark, S., Camargo Jr, C.A.: Mortality after an emergency department visit for exacerbation of chronic obstructive pulmonary disease. COPD: Journal of Chronic Obstructive Pulmonary Disease **3**(2), 75–81 (2006)

[18] Langsetmo, L., Platt, R.W., Ernst, P., Bourbeau, J.: Underreporting exacerbation of chronic obstructive pulmonary disease in a longitudinal cohort. American journal of respiratory and critical care medicine **177**(4), 396–401 (2008)
[19] Leidy, N.K., Wilcox, T.K., Jones, P.W., Roberts, L., Powers, J.H., Sethi, S.: Standardizing measurement of chronic obstructive pulmonary disease exacerbations: reliability and validity of a patient-reported diary. American journal of respiratory and critical care medicine **183**(3), 323–329 (2011)
[20] Leidy, N.K., Wilcox, T.K., Jones, P.W., Murray, L., Winnette, R., Howard, K., Petrillo, J., Powers, J., Sethi, S., Group, E.P.S., et al.: Development of the exacerbations of chronic obstructive pulmonary disease tool (exact): a patient-reported outcome (pro) measure. Value in health **13**(8), 965–975 (2010)
[21] Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)
[22] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[23] McGhan, R., Radcliff, T., Fish, R., Sutherland, E.R., Welsh, C., Make, B.: Predictors of rehospitalization and death after a severe exacerbation of copd. Chest **132**(6), 1748–1755 (2007)
[24] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
[25] Moy, E., Garcia, M.C., Bastian, B., Rossen, L.M., Ingram, D.D., Faul, M., Massetti, G.M., Thomas, C.C., Hong, Y., Yoon, P.W., et al.: Leading causes of death in nonmetropolitan and metropolitan areas—united states, 1999–2014. MMWR Surveillance Summaries **66**(1), 1 (2017)
[26] Quint, J.K., Baghai-Ravary, R., Donaldson, G.C., Wedzicha, J.: Relationship between depression and exacerbations in copd. European respiratory journal **32**(1), 53–60 (2008)
[27] Raju, S., Brigham, E.P., Paulin, L.M., Putcha, N., Balasubramanian, A., Hansel, N.N., McCormack, M.C.: The burden of rural chronic obstructive pulmonary disease: analyses from the national health and nutrition examination survey. American journal of respiratory and critical care medicine **201**(4), 488–491 (2020)
[28] Raju, S., Keet, C.A., Paulin, L.M., Matsui, E.C., Peng, R.D., Hansel, N.N., McCormack, M.C.: Rural residence and poverty are independent risk factors for chronic obstructive pulmonary disease in the united states. American journal of respiratory and critical care medicine **199**(8), 961–969 (2019)
[29] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
[30] Sanchez-Morillo, D., Fernandez-Granero, M.A., Jiménez, A.L.: Detecting copd exacerbations early using daily telemonitoring of symptoms and k-means clustering: a pilot study. Medical & biological engineering & computing **53**, 441–451 (2015)

[31] Seemungal, T.A., Donaldson, G.C., Paul, E.A., Bestall, J.C., Jeffries, D.J., Wedzicha, J.A.: Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine **157**(5), 1418–1422 (1998)

[32] Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum learning: A survey. International Journal of Computer Vision **130**(6), 1526–1565 (2022)

[33] Wilkinson, T.M., Donaldson, G.C., Hurst, J.R., Seemungal, T.A., Wedzicha, J.A.: Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine **169**(12), 1298–1303 (2004)

[34] Xu, W., Collet, J.P., Shapiro, S., Lin, Y., Yang, T., Wang, C., Bourbeau, J.: Negative impacts of unreported copd exacerbations on health-related quality of life at 1 year. European Respiratory Journal **35**(5), 1022–1030 (2010)