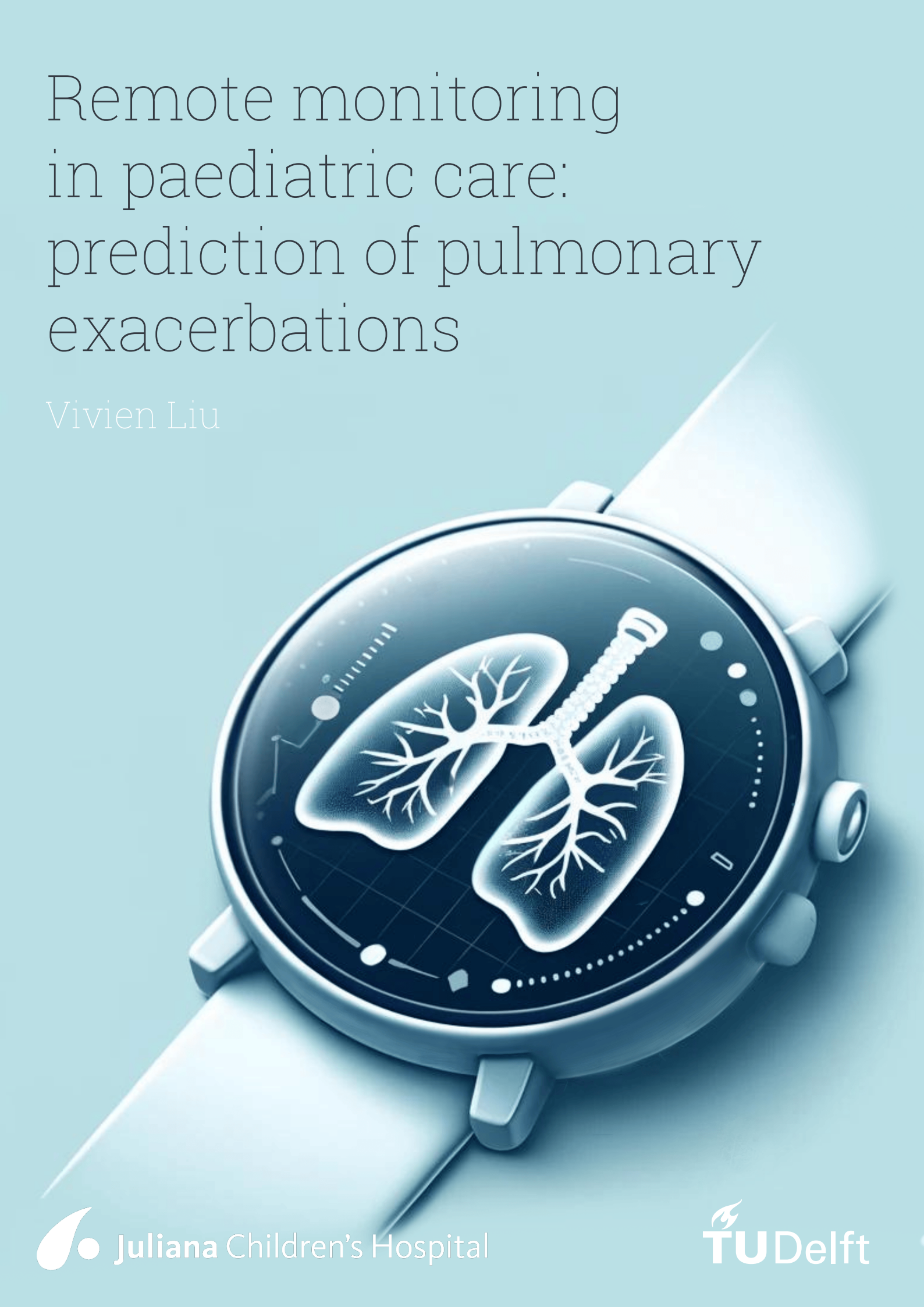
Remote monitoring in paediatric care: prediction of pulmonary exacerbations

Vivien Liu



Juliana Children's Hospital

TU Delft

Remote monitoring in paediatric care: prediction of pulmonary exacerbations

Vivien Liu

Student number: 4646606

13 August 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of General Paediatrics,
Juliana Children's Hospital

February 2024 – July 2024

Supervisor(s):

dr. Matthijs D. Kruizinga
dr. David M.J. Tax
ir. Arman Naseri Jahfari

Thesis committee members:

dr. Niels van der Gaag, MD, PhD, Haga Teaching Hospital (Chair)
dr. Matthijs D. Kruizinga, MD, PhD, Juliana Children's Hospital
dr. David M.J. Tax, Assistant Professor, TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus University Rotterdam

Acknowledgements

This final thesis report concludes my journey in the Master's program in Technical Medicine. In this project, I focused on predicting pulmonary exacerbations in children with asthma and cystic fibrosis using wearable technology in remote patient monitoring. The department of general paediatrics at the Juliana Children's Hospital provided an excellent environment for me to further develop my interest in pediatric care.

I would like to sincerely thank Matthijs for his guidance and support throughout my thesis. Your medical and research expertise were invaluable and greatly helped me navigate the clinical aspects of conducting patient research. I also want to express my gratitude to David for your enthusiasm and patience in explaining key concepts in machine learning. Your insights made this challenging topic more approachable and enjoyable.

I am also grateful to the research group at the Juliana Children's Hospital for their continuous support and positivity. I hope the research room continues to flourish, both with new research ideas and the greenery of its ever-growing collection of plants.

Finally, I want to extend a special thank you to Arman. Our coffee breaks and conversations provided me with much-needed motivation and insight. Whether addressing small or large tasks, your willingness to help and explain encouraged me greatly throughout this graduation project. Your perspective and openness on various topics gave me new insights, and your support played a key role in helping me complete this project.

Vivien Liu
The Hague, August 2024

Abstract

Introduction

Pulmonary exacerbations are critical events in paediatric patients with asthma or cystic fibrosis (CF). These exacerbation events are often associated with sudden health deterioration and increased healthcare burden. The early prediction of exacerbations events could allow for timely interventions, and thus improved patient outcomes. This thesis attempted to develop a machine learning (ML) model to predict pulmonary exacerbations before they occur in a paediatric population using remote patient monitoring (RPM) data.

Methods

A retrospective study was conducted using continuous data from wearable devices, daily spirometry, environmental data, and patient-reported outcomes. Predictions were focused on the occurrence of an exacerbation within three prediction windows (1-day, 3-day, and 7-day). Two ML approaches were considered: anomaly detection (using Gaussian mixture model, Isolation forest, One-class-SVM, and Local outlier factor), and classification models (Logistic regression, Random forest), using 5-fold nested cross-validation. Time-related transformations were performed to capture the temporal dependency of time-series data, including the feature engineering of clinical features related to heart rate and physical activity.

Results

A total of 2401 home monitoring days of 90 paediatric patients, with 10 observed exacerbation events were included in the analysis. All models struggled to achieve high predictive value, with PR-AUC values below 0.20 and ROC-AUC values ranging from 0.43 to 0.72 across different time windows. No single model consistently outperformed the others. Despite the low performance, the models demonstrated better than random prediction for secondary outcomes, such as weekends and holidays, suggesting the ability to capture patterns in the data.

Conclusion

This thesis shows the potential and limitations of using ML techniques for predicting pulmonary exacerbations using RPM data. The current anomaly detection and classification model performances are insufficient for clinical application. The low incidence of exacerbation events and the limitations in data quality contribute to these results. These findings point to the need for further refinement and more robust datasets to fully realise the potential of ML in the context of predicting pulmonary exacerbations.

Nomenclature

Abbreviation	Definition
ACD-6	Asthma Control Diary, six-question version
ACQ	Asthma Control Questionnaire
AUC	Area Under the Curve
BMI	Body Mass Index
CF	Cystic Fibrosis
CFQ-R	Cystic Fibrosis Questionnaire-Revised
CV	Cross-Validation
FEV1	Forced Expiratory Volume in 1 second
FPR	False Positive Rate
FVC	Forced Vital Capacity
GMM	Gaussian Mixture Model
ICS	Inhaled Corticosteroids
KNN	k-Nearest Neighbors
LABA	Long-Acting β-Agonist
LOF	Local Outlier Factor
LR	Logistic Regression
LSTMs	Long Short-Term Memory networks
ML	Machine Learning
nCV	Nested Cross-Validation
OC-SVM	One-Class Support Vector Machine
PAQLQ	Paediatric Asthma Quality of Life Questionnaire
PCA	Principal Component Analysis
PEF	Peak Expiratory Flow
PR-AUC	Precision-Recall Area Under the Curve
PedsQL	Pediatric Quality of Life Inventory
RNN	Recurrent Neural Networks
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
RPM	Remote Patient Monitoring
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine

Contents

1

Introduction

1.1. Asthma and cystic fibrosis

Asthma and cystic fibrosis (CF) are prevalent chronic pulmonary diseases in the paediatric population. In 2021, approximately 7% of the children in the Netherlands between ages 7 and 20 were experiencing asthma, with an even larger prevalence observed in younger children [1]. Paediatric asthma is characterised by chronic airway inflammation with variable expiratory airflow obstruction [2, 3]. Common symptoms of asthma include coughing, wheezing, dyspnea, chest tightness triggered by physical activity, respiratory infection, and allergies, which influence the patient's quality of life.

It is understood that asthma is a multifactorial disease, caused by many environmental factors such as exposure to allergens, air pollutants, irritants, and cigarette smoke, and by genetic factors. Studies with twins have shown that asthma has a strong genetic component and tends to run in families [4]. Asthma onset can occur at any point in life, and some children who experience asthma during childhood will continue to have the condition into adulthood [3]. According to the most recent CF registration of 2022, a total of 548 children are currently receiving treatment in Dutch children's centers [5].

Paediatric cystic fibrosis is one of the most frequently diagnosed hereditary diseases in the Western population, affecting different organs such as the lungs, intestine, and pancreas [6, 7]. The prevalence of CF varies worldwide, but Europe is amongst the continents with the highest prevalence [7]. A large spectrum of clinical presentation of CF is present, including lung disease, pancreatic disease, liver disease, chronic pansinusitis, nasal polyposis, and elevated salt levels in sweat gland secretion. CF itself is caused by mutations in a gene called cystic fibrosis transmembrane conductance regulator (CFTR) [7, 8]. There have currently been over 2000 distinct mutations identified as the cause of the disease. These CFTR mutations have varying effects on CFTR protein synthesis, functionality, and stability at the cell membrane.

Asthma is mainly caused by genetic and environmental factors that result in persistent inflammation of the airways, while cystic fibrosis (CF) is a hereditary condition that affects several organs because of faulty CFTR proteins. Despite this, their clinical presentations are comparable. For instance, cough, dyspnea, and exacerbations are pulmonary symptoms that are common to both diseases.

1.2. Pulmonary exacerbations

Pulmonary exacerbations are episodes of acute worsening of respiratory symptoms from the patient's baseline, requiring additional treatment or hospitalisation [9, 10]. Common triggers for exacerbations in children include (viral) upper airway infections [11] or environmental exposures [12]. These exacerbations can affect lung disease progression and the quality of life of children and their families. Exacerbations can result in the children and their parents missing school and work days; limit the child's physical and social activities; require additional health-related costs; and lead to emergency department visits or hospital admissions [13–15].

There is currently no universally accepted definition for pulmonary exacerbations [16]. Pulmonary exacerbations are patient-specific and they are often presented by a variety of symptoms such as increased cough, dyspnea, increased sputum production, deterioration of lung function parameters, weight loss, and decreased energy level and appetite [17, 18]. Despite this lack of a clear definition, the general agreement is that exacerbations are (sub)acute deteriorations in symptom control, sufficient to cause risk to health, and require a change in treatment [19].

The management of chronic pulmonary diseases focuses on the importance of maintaining symptom control and reducing the risk of future exacerbations through monitoring and risk assessment [2, 7]. Symptoms can be reduced using treatments such as anti-inflammatory drugs, bronchodilators, and biologicals, along with non-pharmacological measures such as environmental control and patient education. Children, however, may find it more difficult to recognise symptoms of exacerbation and address it to their parents. This difficulty in recognising and managing symptoms can lead to delays in seeking appropriate medical help [20]. Therefore, regular follow-up is needed to prevent disease deterioration and future exacerbations [21].

1.3. Remote Patient Monitoring

Remote patient monitoring (RPM) is a form of telehealth that uses technology to monitor patients' health status outside of the hospital setting [22]. RPM has been a rapidly growing industry over the past century, particularly since the COVID-19 pandemic [23, 24]. The use of RPM offers several benefits, such as early detection of health deterioration, which leads to timely interventions, improved patient outcomes, and reduced healthcare costs [25].

The introduction of personal smart devices has allowed for real-time monitoring of physiologic parameters such as heart rate, activity levels, oxygen saturation, respiratory rate, sleep patterns, and GPS location [26–28]. For pulmonary diseases, such devices may provide valuable insight

into disease control at home and the risk of exacerbations. Although existing literature regarding monitoring technologies in asthma or CF is heterogeneous, these tools have shown promising results for RPM in paediatric care [29].

In particular, studies with wearables have shown the potential to predict exacerbations [30–32]. These tools are passive monitoring tools that collect patient data with minimum active user engagement. Therefore, wearables may reduce the impact of exacerbations through early recognition of symptoms and timely treatment. As large amount of data is gathered through RPM, it is common to utilise machine learning (ML) techniques to develop their prediction models [33]. ML algorithms analyse large data sets collected from wearables to identify patterns and correlations that might not be evident from traditional statistical methods. By training these models on historical data, they can learn to predict future exacerbations based on subtle changes in monitored parameters.

Although these potential benefits are extremely valuable, RPM evidence and its use for paediatric pulmonary disease management are still in their early stages. Many studies focus on assessing the relation of home-monitoring parameters with varying indices (such as symptom scores, wheezing, and exacerbations). However, existing research seems to lack multiparameter RPM methods combined with ML for predicting future exacerbations in pediatric healthcare. [34]

1.4. Thesis objective

This thesis focuses on the development of a machine learning algorithm to predict pulmonary exacerbations in children with asthma or cystic fibrosis within 7 days of onset, based on remote patient monitoring data. A secondary aim of this study was to predict additional outcome measures, such as the heightened symptom days of patients with asthma and cystic fibrosis indicated by the clinical questionnaire scores, using the home monitoring dataset.

2

Data collection and preparation

2.1. Study population

For this thesis, a retrospective study was conducted with data acquired from a clinical validation study with paediatric patients aged 6–16 years diagnosed with either asthma or CF [35]. The clinical study aimed at the clinical validation of smartwatch biomarkers (physical activity, heart rate, and sleep) and portable spirometer biomarkers (FEV1; forced expiratory volume in 1 second and FVC; forced vital capacity) in children with asthma and CF. Patients in this study were recruited from the outpatient clinic at the Juliana Children's Hospital (Haga Teaching Hospital, The Hague, The Netherlands) and Sophia Children's Hospital (Erasmus Medical Centre, Rotterdam, The Netherlands), and the study was conducted between November 2018 and February 2020. The diagnosis of asthma was based on clinical symptoms combined with pulmonary function tests (PFTs), and the diagnosis of CF was confirmed using genetic tests.

2.2. Data acquisition

The data from asthma and CF patients was collected using various digital devices, comprising a combination of wearable, spirometry, environmental, and clinical questionnaire data. All patients were monitored over a total duration of 28 days, and an overview of all available features of the dataset is shown in Table 2.1

Each patient wore a Steel HR smartwatch, which continuously measured the physical activity (amount of steps taken) through an accelerometer and the heart rate via a photoplethysmography sensor. Several sleep-related parameters were also calculated automatically by the smartwatch, such as the average heart rate during sleep and the wake-up count. Furthermore, patients performed daily home-based spirometry using the Air Next spirometry device. This spirometer measured the FEV1, FVC, and peak expiratory flow (PEF).

Clinical characteristics were collected through questionnaires. Daily questionnaires included

the six-question Asthma Control Diary (ACD-6) for asthma patients and a daily respiratory symptom questionnaire for CF patients. For the clinical baseline characteristics, parents filled out the Pediatric Quality of Life Inventory (PedsQL 4.0) questionnaire, children with asthma the Asthma Control Questionnaire (ACQ) and the Paediatric Asthma Quality of Life Questionnaire (PAQLQ), and children with CF filled out the Cystic Fibrosis Questionnaire-Revised (CFQ-R). External data was obtained from the electronic patient files (e.g., prescribed medications) and the Royal Dutch Meteorological Institute (e.g., amount of pollen in the air).

Table 2.1: Overview of features (n=194) in the acquired dataset. Wearable features were recorded continuously, while spirometry, questionnaires, and environmental data were gathered daily. Baseline characteristics and specific questionnaires (PedsQL 4.0, ACQ, PAQLQ, and CFQ-R) were assessed once at the beginning of the study. The dataset includes patients monitored over a 28-day study period.

Features (n=194)	Description
Wearable (n=69)	
Daily Activity (n=29)	*Total number of steps over 24 hours, sorted into hourly values.* *Maximum number of steps during the most active hour.* *The number of steps taken between 15:00 and 19:00.*
Heart rate (n=32)	*Average heart rate over 24 hours, sorted into hourly values.* *Average, minimum, maximum HR during awake and sleep period.* *5th and 95th percentile of all heart rates measured during a day.*
Sleep (n=8)	*Awake, light, deep, total sleep duration in seconds. Sleep scores.* *Number of times woken up, sleep and wake time.*
Spirometry (n=13)	
Lung function (n=13)	*Measured and predicted amount of air expired in 1 second (FEV1).* *Measured and predicted total amount of air expired (FVC).* *Calculated and predicted ratio between FEV1/FVC.* *Measured and predicted peak flow (PEF).* *Assessment of technique performance, graded spirometry curves.*
Questionnaire (n=39)	
Questionnaire (n=39)	*ACD-6 scores, daily respiratory symptom questionnaire scores.* *PedsQL 4.0 scores, ACQ scores, PAQLQ scores, and CFQ-R scores.*
Other (n=73)	
Baseline characteristics (n=45)	*Age, gender, height, weight, BMI, school year, race, sports, pets.* Age disease diagnosis, asthma family history, *smoking situation.Type urbanisation, activity scores, medication use, clinical condition.*
Environmental Data (n=19)	*Amount of pollen in air, pollutant concentrations.* *Wind speed, temperature, rainfall, sunshine duration.*
Miscellaneous (n=8)	*Subject number, school day, day type, screen time.* *Day number, month, weekday, week.*
Exacerbation (n=1)	*Onset of exacerbation (day 0).*

ACD-6: six-question Asthma Control Diary; PedsQL: Pediatric Quality of Life Inventory; ACQ: Asthma Control Questionnaire; PAQLQ: Paediatric Asthma Quality of Life Questionnaire; CFQ-R: Cystic Fibrosis Questionnaire-Revised; BMI: body mass index.

In the available dataset, the continuously measured parameters (heart rate and number of steps) were available in hourly intervals. For example, the number of steps measured between 15:00 and 16:00 or the average heart rate within the same hourly time frame. Other daily measurements were either measured (e.g., best FEV1) or calculated (e.g., the average heart rate during

sleep). There was no specification of when these daily measurements were taken.
Certain features only had one value for the entire patient monitoring period, as these were baseline parameters composed of the patient characteristics and clinical questionnaire answers.

2.2.1. Outcome variable

For the primary analysis, the outcome variable was defined as the occurrence of an exacerbation within a certain number of days. An exacerbation was defined differently for asthma and CF. For asthma, an exacerbation was identified as worsening of symptoms requiring the use of systemic corticosteroids to prevent a serious outcome [17]. In the case of CF, this was defined as the need for additional antibiotic treatment due to a recent change in symptoms or decrease in pulmonary function (≥10% of predicted FEV1) [36]. To predict the occurrence of pulmonary exacerbations within an upcoming period, various prediction windows were considered. Specifically, three different prediction windows were analysed: a 1-day window, a 3-day window, and a 7-day window. For instance, a prediction within the 3-day window would correspond to an exacerbation occurring within the next three days. Table 2.2 provides a schematic overview of the 1-day, 3-day, and 7-day prediction windows, indicating the period during which the model anticipates the onset of an exacerbation.

Day	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Onset of exacerbation															
1-day window															
3-day window															
7-day window															

Table 2.2: Three variations of the exacerbation time window (1-day, 3-day, and 7-day). The onset of an exacerbation is considered as day 0.

2.3. Data preprocessing

2.3.1. Data exclusion

Patient days with a watch wear time below 50% between 06:00 and 22:00 were removed due to data insufficiency. Additionally, only spirometry sessions with at least one acceptable spirometry measurement (graded A-E according to the American Thoracic Society/European Respiratory Society criteria [37]), were included for the analysis. Given, that the dataset was comprised of two different patient populations with their own unique features (e.g., ACQ, PAQLQ, CFQ-R, the use of inhalation medicine for asthma, and the presence of pancreatic insufficiency in CF), features specific to one population were excluded in the primary analysis to maintain consistency. Lastly, features missing data for the entire 28-day study period for any patient were also removed.

2.3.2. Data transformation

All categorical features (e.g., gender, screen time) were transformed from strings into numeric formats using scikit-learn's *LabelEncoder* [38] for the effective use of ML algorithms.
Patients were monitored over multiple days and their respective data was normalised and imputed separately. This allowed for adjustments based on each patient's baseline and variability, as opposed to performing these transformations across the whole dataset. By applying scikit-learn's *StandardScaler* [39], the patient data was standardised to have a mean of zero and a standard deviation of one. Missing values in the dataset (NaN, Not a Number) were filled using k-Nearest Neighbours (KNN) imputation [40]. A method that replaces missing values with the average of the K-amount of nearest neighbours in the dataset.

2.3.3. Feature engineering

In order to further enhance the predictive capability of the clinical time-series data, new features were engineered. These included clinical features related to heart rate and physical activity, as well as time-series transformations designed to capture temporal dynamics.

Three heart rate features were developed: The first feature, the nocturnal heart rate reserve, represents the difference between the resting heart rate and the maximum heart rate during sleep [30]. The second feature captures the difference between measured heart rates and age-specific normal pediatric heart rates, according to Fleming et al. [41]. This feature included the differences of the daily average heart rate, hourly heart rates, as well as the 95th and 5th percentile heart rates. The third clinical feature was developed to capture the relationship between heart rate and physical activity levels [42]. For further explanation and the specific formulas used to derive these features, see Appendix A.1.
In addition to the clinical features, time-series transformations were performed. Rolling windows were utilised to calculate statistics (mean, standard deviation, maximum, and minimum) over a fixed time interval [43]. For example, a 7-day window aggregated data from the current day and the six preceding days. Moreover, a lagged variable was used to account for the influence of previous time steps. This feature represented the data observed one day prior. Finally, the first differences were calculated by measuring the change between consecutive data points. This feature captured the shift from one day to the next.

For the prediction of pulmonary exacerbations, the following features were engineered: (a) 3-day mean, standard deviation, maximum and minimum, (b) 5-day mean, standard deviation, maximum and minimum, (c) 1-day lag, and (d) first differences of the daily heart rate (awake and during sleep), daily step rate, best measured lung function parameters, environmental parameters, and the new engineered clinical features.

2.4. Model development

For model development, both anomaly detection models and classification models were used to predict outliers indicating exacerbations. This included the *Gaussian Mixture Model*, *Isolation*

Forest, *One-Class Support Vector Machine*, and *Local Outlier Factor* for anomaly detection, and *Logistic Regression* and *Random Forest* for classification. All models were implemented using the scikit-learn library.

2.4.1. Anomaly detection

The principal task of anomaly detection is to identify data samples that do not fit the overall distribution. See Figure 2.1 for a schematic representation of the anomaly detection models.

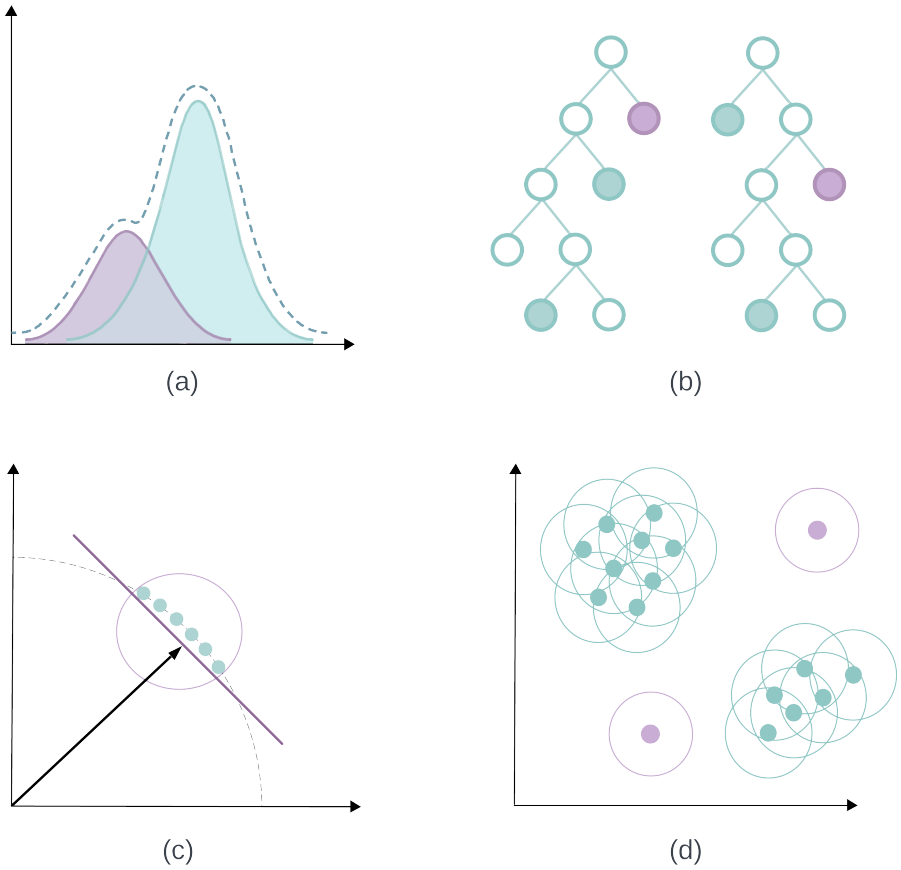


Figure 2.1: Schematic representation of the anomaly detection models: (a) Gaussian Mixture Model, (b) Isolation Forest, (c) One-class Support Vector Machine, and (d) Local Outlier Factor.

Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a probabilistic model, that models the data as a mixture of Gaussian distributions [44, 45]. In Figure 2.1a, each distribution represents a cluster in the dataset, and the model is a weighted sum of these distributions. The goal is to find parameters of distributions which describe the samples the best. The GMM calculates the probability of a sample belonging to each cluster. Samples with a low probability belonging to any cluster are identified as anomalies.

Isolation Forest

Isolation Forest (IF) is an anomaly detection method that isolates observations by randomly selecting a feature and then randomly selecting a split value within the range of the maximum

and minimum values of the selected feature [46, 47]. In Figure 2.1b, this process is represented by decision trees, where each tree isolates a subset of the data. IF is based on the assumption that anomalies are few and different, therefore anomalies are more susceptible to a mechanism called isolation. Only a limited number of conditions are required to separate the anomaly cases, making the isolation of anomalies easier. On the contrary, isolating normal observations requires more conditions. An anomaly score can be calculated from the number of conditions required to isolate a given observation. Specifically, the anomaly score is calculated based on the path length from the root to the leaf for each observation. Anomalies tend to have shorter path lengths and normal observations have longer path lengths.

One-Class Support Vector Machine
One-Class Support Vector Machine (SVM) aims to learn a boundary or decision function that best separates normal data from anomalies in a transformed high-dimensional space [48]. One-Class SVM exclusively trains on data points from the normal behavior of the data. During training, a binary function that identifies whether new data instances belong to the normal class or anomalies is derived. This boundary is depicted in Figure 2.1c, and is shown as a line (hyperplane) or a sphere (hypersphere). The algorithm separates all data instances in a feature space from the origin, and then maximises the distance from the origin to the separating boundary. The result is a function that classifies instances as normal if they are inside the separating boundary whereas the observations outside the boundary are predicted as anomalies.

Local Outlier Factor
Local Outlier Factor (LOF) is an anomaly detection algorithm based on the concept of local densities [49]. In Figure 2.1d, the data points are shown with surrounding circles representing their local density. LOF measures the local deviation of density of a given sample with respect to its neighbours. A score is assigned to the sample, which is used as a measure of the 'deviation degree'. The larger the deviation, the larger the LOF score. Anomalies are considered to have relatively smaller densities and therefore have larger LOF scores.

2.4.2. Classification models

For classification models, the goal is to assign labels to unseen data samples, based on patterns and relationships learned from the training data. See Figure 2.2 for the schematic representations of the classification models.

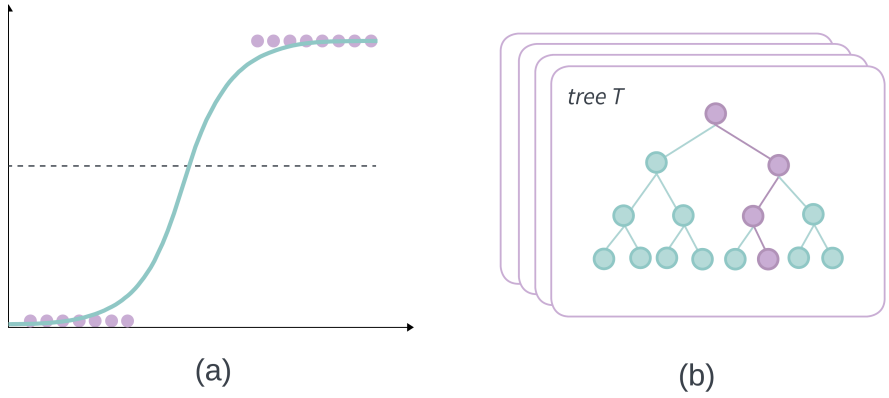


Figure 2.2: Schematic representation of the classification models: (a) Logistic Regression and (b) Random Forest.

Logistic regression

Logistic regression (LR) is a commonly used classical statistical model, for the probability estimation of a binary event occurring [50]. In Figure 2.2a, the LR model is illustrated by an S-shaped curve, known as the Sigmoid function, which maps any input value to a probability between 0 and 1. The data points on either end of the curve represent the two possible outcomes, while the dashed line in the middle represents the decision boundary—the threshold at which the model decides between the two outcomes. LR is a probability-based classification algorithm, which assumes a linear relationship with the logit (natural logarithm of the probabilities) of the outcome, introducing a non-linearity in the form of the Sigmoid curve. A limitation of LR, however, is the assumption of linearity between the dependent and independent variables [51]. This might restrict the level of complexity needed to adequately model certain prediction problems, in which ML methods may lead to better results.

Random Forest

Random forest (RF) is an ensemble learning technique, which uses both bagging and decision tree concepts [52]. The bagging method refers to generating a new dataset with replacement from an existing dataset. This creates diverse training sets, which are used to train different models in the ensemble. Decision trees have a flowchart-like structure, with nodes splitting the data set into smaller subgroups based on the input feature. For the construction of the RF, multiple decision trees (forest) are built using randomly selected training datasets and subsets of the predictor values. Figure 2.2b depicts this concept by showing several decision trees, each with its own unique path leading to a particular outcome. The results from each tree are then aggregated to give a prediction for each observation.

2.4.3. Splitting and hyperparameter tuning

Cross-validation (CV) is a common method to assess the performance of ML models by optimising the split of the dataset. CV involves dividing the dataset into multiple training and validation sets. The evaluation is repeated across multiple validation sets, to estimate the generalisation performance of the model. However, the disadvantage of CV is its possibility of overfitting, as the same data is used for both tuning and evaluation. To reduce this risk of overfitting, nested

cross-validation (nCV) can be used. nCV involves an outer CV layer for model performance assessment and an inner CV layer for hyperparameter tuning, reducing the possibility of bias in the model's performance. For this study, nCV was applied with a 5-fold split and stratified group CV, in both the inner and outer folds. This approach ensured that the data from any given subject was only included in either the training or validation/test sets within each inner and outer CV loop, while also maintaining the proportion of the data.

In the case of anomaly detection, novelty detection was used in combination with nCV. Novelty detection is particularly useful for extremely imbalanced classes [53]. The objective of novelty detection is to determine whether an instance belongs to the 'normal' class [54]. During the training phase, the models are only trained on normal data. In the evaluation phase, data with both inliers and outliers are included. The model then assigns a novelty score to each instance, and it is expected that the model assigns significantly different novelty scores to outliers compared to inliers.

The nCV strategies for both anomaly detection and classification are visually represented in Figure 2.3.

Figure 2.3: Visual representation of the 5-fold nested cross-validation strategy for (a) novelty detection and (b) classification. For novelty detection, the algorithm trains on only the inliers (purple), and for classification, both the inliers and outliers are used for training. Although not explicitly represented for clarity, each fold contains data of a unique subject.

The inner layer on the nCV was used for hyperparameter tuning. The hyperparameter settings for the different models are displayed in Appendix A.2. These settings were often combined with

the tuning of the number of principal components for the principal component analysis (PCA). This is an unsupervised algorithm, commonly used for dimensionality reduction, in which new variables are computed as linear combinations of the original features [55].
The hyperparameter tuning was optimised based on maximising the area under the precision-recall curve (PR-AUC). This performance metric was assessed over all possible threshold values, which allowed the model to measure the trade-off between precision and recall. Therefore, the AUC-PR is particularly useful for imbalanced class problems, as it focuses on identifying the minority class. Moreover, the combination of hyperparameters leading to the best performance measure was used for the outer test fold.

2.4.4. Model evaluation

In the outer folds, the trained models were finally evaluated on the test sets, which contained new and unseen data. This cross-validation process was repeated five times resulting in five test sets and the final performance was then composed of the average performances of the five outer folds.
The full evaluation metrics for the test sets included the following performance measures, the PR-curve (with PR-AUC), the receiver operating characteristic curve (ROC) including its area under the curve (ROC-AUC), precision, recall (sensitivity), specificity, and the F1-score. See appendix A.3 for a comprehensive description of the performance metrics. In addition, the feature importance scores for the random forest model were obtained, with the 20 features highest in importance.

2.4.5. Baseline performance

The baseline comprises a synthetic normalised dataset consisting of inliers and outliers with a similar class imbalance distribution was created. This synthetic dataset was used to establish a baseline reference for model performance assessment. The objective was to understand the model's predictive capability of the anomalies, by changing the characteristics of these outliers.

The synthetic normalised dataset consisted of features with a mean of zero and a standard deviation of one. For the calculation of the baseline performance, the mean of the anomalies was shifted in small increments. Each mean shift represents a deviation of the anomaly data values, with respect to the normal data values. For example, an anomaly mean shift of 2.0 corresponds to anomalies that are (on average) two standard deviations away from the mean of the normal data. At each mean shift, the AUC-PR was evaluated and recorded. These results were plotted and displayed as a visual change of the performance metric over a mean shift of the anomalies.

2.5. Secondary Analysis

In addition to the primary analysis for the prediction of exacerbation days, a secondary analysis was performed to explore additional outcomes, such as the prediction of heightened symptom days prior to one day before they occurred. The dataset consisted of the wearable, spirometry,

environmental, and patient characteristic data. For asthma patients, a symptom day was defined by an ACD-6 score of 1.5 or higher. For CF patients, a CF symptom score of 7 or higher was considered a symptom day. Given that questionnaire results are subjective and heightened symptom days can be perceived earlier than when the aforementioned threshold values are reached, the performances were also evaluated for different thresholds for the ACD-6 and CF symptom scores.

For this analysis, the individual asthma and CF population datasets were used for the prediction of the heightened symptom days. The patient population-specific features, which were previously excluded (Section 2.3.1), were now included in this analysis. Therefore, except for the patient dataset and outcome variables, the methods for data transformation, feature engineering, classifier algorithms, splitting, hyperparameter tuning, and model evaluation were consistent with those used in the primary analysis.

3

Results

3.1. Study cohort

The study cohort consisted of a total of 90 patients, monitored from November 2018 to February 2020. Characteristics of the study cohort are presented in Table 3.1.

Table 3.1: Characteristics of the study cohort

Characteristics	Asthma	Cystic Fibrosis	Total
Patient, N	60	30	90
Total amount of measured days, N	1570	831	2401
Age, median (25-75)	11 (8-12)	10 (7-12)	10 (8-12)
Gender (male), N (%)	40 (66.7%)	14 (46.7%)	54 (60%)
BMI, mean (std)	19.3 (4.2)	16.3 (1.6)	18.3 (3.9)
LABA therapy, N (%)	35 (58.3%)		
ICS, N (%)	58 (96.7%)		
Pancreas insufficiency, N (%)		28 (93.3%)	
Wearable parameters			
Daily heart rate (awake), mean (std)	87.3 (9.7)	85.3 (9.8)	86.6 (7.8)
Daily heart rate (sleep), mean (std)	73.5 (8.6)	71.5 (8.6)	72.8 (8.7)
Sleep duration, mean (std)	8.7 (1.3)	9.2 (1.4)	8.9 (1.3)
Daily steps, mean (std)	6516.6 (3684.0)	6752.0 (3119.7)	6597.3 (3502.1)
Spirometry parameters			
FEV1, mean (std)	2.0 (0.8)	1.8 (0.6)	2.0 (0.7)
FVC, mean (std)	2.7 (1.0)	2.2 (0.7)	2.5 (0.9)
PEF, mean (std)	4.2 (1.7)	3.8 (1.4)	4.1 (1.6)
Outcomes			
Exacerbations, N (%)	5 (0.3%)	5 (0.6%)	10 (0.4%)
ACD-6, mean (std)	0.8 (0.9)		
Respiratory symptom score, mean (std)		4.7 (4.1)	

Descriptive statistics of the study cohort. BMI: body mass index; LABA: long-acting β-agonist; ICS: inhaled corticosteroid; FEV1: forced expiratory volume in 1 second; FVC: forced vital capacity; PEF: peak expiratory flow; ACD-6: six-question Asthma Control Diary; std: standard deviation.

The dataset included 60 children with asthma and 30 children with CF, with ages ranging from 6 to 16 years old (median (25-75): asthma 11 years (8-12), CF 10 years (7-12)). A total of 2612 patient days were monitored. During this monitoring period, 11 exacerbation events were observed (0.5% of total daily measurements) with five asthma, and six CF exacerbations. However, one patient did not wear the smartwatch during the onset of the event, reducing the total number of recorded events to 10. After the removal of patient data days due to insufficient watch wear time, 2401 patient data days were left, representing a removal of 8.1%. Following this, adherence to daily spirometry measurements was 72% across the entire dataset. Out of these measurements, 87.5% (986 out of 1127) were of adequate quality (graded A-E), while the remaining 12.5% (graded F or U) of the lung function measurements were excluded.
Figure 3.1 illustrates the time series for the average heart rate, number of steps taken, and best FEV1 for a patient without and with exacerbation. Empty gaps in the time series indicated missing values in the dataset. The dataset had varying levels of missing data across different categories of features. Approximately 6.6% of the wearable features had missing values, with 30.6% of the spirometry features missing values, and 24.7% of the questionnaire-based features had incomplete data.

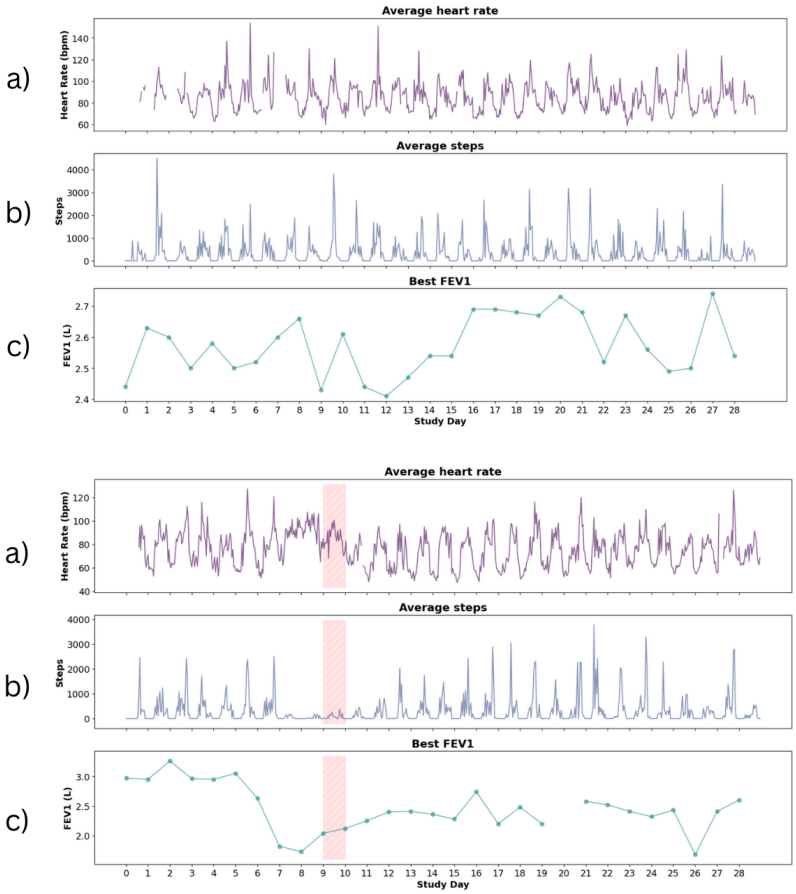


Figure 3.1: Time series for patients without (top) and with (bottom) an exacerbation. (a) Average heart rate (bpm), (b) average steps, and (c) Best FEV1 (L). The onset of an exacerbation is marked as a red dashed area.

3.2. Baseline performance

A synthetic normally distributed dataset with anomalies was constructed to evaluate the performance of the models. This dataset was composed of 2000 samples with 125 features and a skewed distribution. In this dataset, 0.4% of these samples were defined as anomalies. The performance was calculated using this synthetic dataset while shifting the mean of the anomalies from 0 to 2.3, increasing the distinction between an outlier and an inlier. Figure 3.2 shows the change in performance due to the mean shift of the outliers in anomaly detection models (left) and classification models (right).

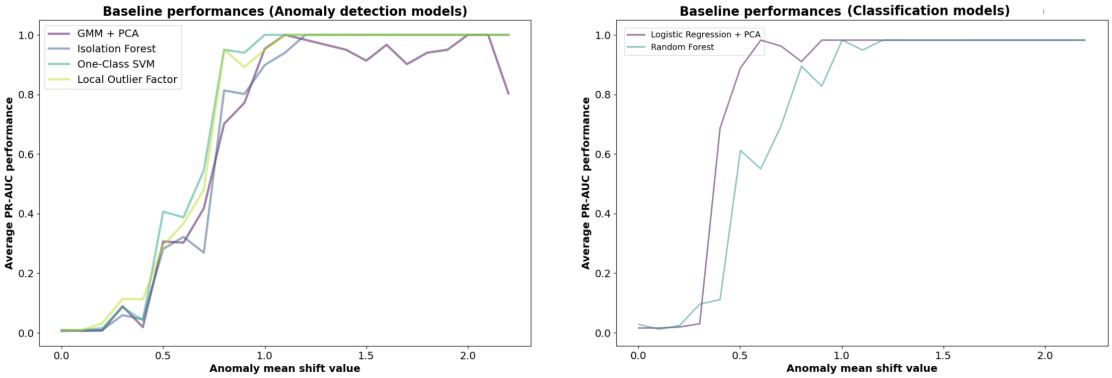


Figure 3.2: Baseline performances of the anomaly detection models (Left: Gaussian Mixture Model, Isolation Forest, One-class SVM, and Local Outlier Factor) and classification models (Right: Logistic Regression and Random Forest) with the shift in mean of the outliers.

In this figure, it can be seen that the performance of the models improved with an increasing mean shift of the anomalies. Considering the anomaly detection models, OC-SVM and LOF demonstrated a slight edge over the other two models in detecting the outliers earlier. They both achieved good performances (>0.8) at mean outlier shifts of 0.75 or higher. For the classification tasks, LR outperformed RF, with good performances acquired at mean outlier shifts of 0.50 or higher. Furthermore, it was observed that the performance of the GMM model became less stable at larger mean shifts, whereas the other anomaly detection models consistently maintained a PR-AUC of 1.0 at higher shift values.

3.3. Primary analysis

3.3.1. Anomaly detection performance

Table 3.2 lists all the average AUCs after 5-fold nCV of the anomaly detection models for both the PR-curve and ROC-curve, based on different window sizes for the outcome variable. The following window sizes were considered: 1-day window, 3-day window, and 7-day window. In addition, Figures 3.3 and 3.4 display the full PR-curves and ROC-curves, respectively, of their corresponding algorithms and outcomes. The complete performance evaluation results can be seen in Appendix A.4, including the tuned hyperparameters.

Figure 3.3 and Table 3.2 illustrate that for each time window, all four anomaly detection models

scored similar and low PR-AUCs of around 0.04. There was no clear superior algorithm across all evaluated results. Additionally, Figure 3.4 and Table 3.2 display moderate performances of the ROC-AUC of approximately 0.60, in combination with standard deviations of around ±0.15. Notably, both the average ROC-AUCS and standard deviations for the 1-day window size were slightly higher compared to the 7-day window outcome.

Table 3.2: Performance evaluation results (PR-AUC and ROC-AUC) of the anomaly detection models for different window sizes outcomes (1-day, 3-day, and 7-day window).

Anomaly Detection						
	1-day window		3-day window		7-day window	
	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC
GMM + PCA	0.03 (0.01)	0.67 (0.20)	0.04 (0.01)	0.60 (0.17)	0.07 (0.04)	0.58 (0.17)
Isolation Forest	0.05 (0.3)	0.62 (0.25)	0.04 (0.02)	0.60 (0.17)	0.05 (0.02)	0.58 (0.18)
One-class SVM	0.04 (0.02)	0.65 (0.25)	0.04 (0.02)	0.60 (0.20)	0.05 (0.02)	0.51 (0.12)
Local Outlier Factor	0.03 (0.01)	0.67 (0.20)	0.02 (0.01)	0.55 (0.06)	0.04 (0.01)	0.45 (0.10)

PR-AUC: precision-recall area under the curve; ROC-AUC: receiver operating characteristic area under the curve; GMM: gaussian mixture model; SVM: support vector machine.

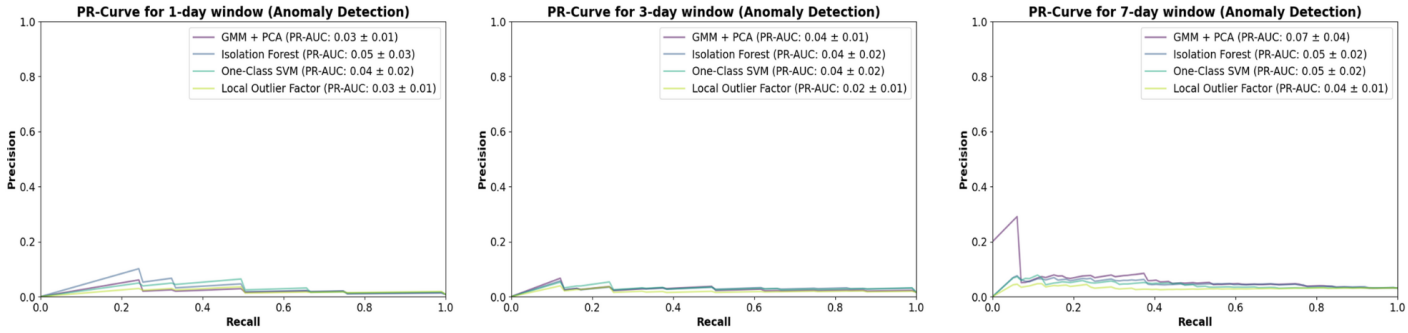


Figure 3.3: Precision-recall curves of the anomaly detection models for a 1-day window (left), 3-day window (middle), and 7-day window (right) as the outcome variable.

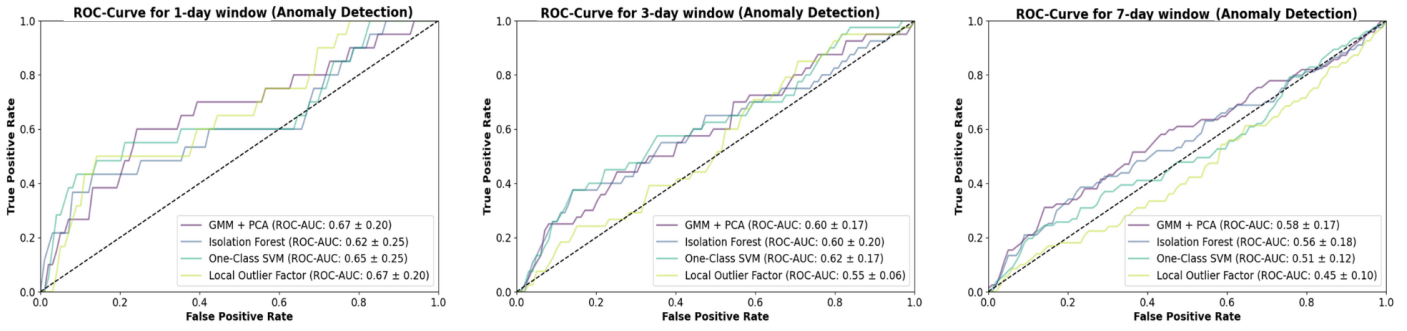


Figure 3.4: Receiver operating characteristic curves of the anomaly detection models for a 1-day window (left), 3-day window (middle), and 7-day window (right) as the outcome variable.

3.3.2. Classification model performance

Similar to the previous subsection, performances were evaluated after 5-fold nCV of the classification models for the different window sizes of the outcome variable. These results are shown in Table 3.3 and Figures 3.5 and 3.6, in which the figures display the full PR-curves and ROC-curves, respectively, for the classification algorithms. Detailed performance evaluation results, including the tuned hyperparameters and feature importance scores for the random forest model, are presented in Appendix A.5.

Figure 3.5 and Table 3.3 show that all classification models scored low PR-AUCs. Logistic regression seemed to perform better compared to random forest, which was consistent with the baseline performance. However, the standard deviation for logistic regression was relatively also larger. Similarly, the ROC-AUCs displayed in Figure 3.6 and Table 3.3, showed moderate results (0.43-0.72), with higher performance results for the 1-day window compared to the 7-day window outcome. Furthermore, considering the 1-day window size, the performance metrics of the classification models were slightly higher compared to those of the anomaly detection models.

Table 3.3: Performance evaluation results (PR-AUC and ROC-AUC) of the classification models for different window sizes outcomes (1-day, 3-day, and 7-day window).

Classification Model						
	1-day window		3-day window		7-day window	
	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC
Logistic Regression	0.15 (0.18)	0.72 (0.21)	0.08 (0.09)	0.43 (0.25)	0.05 (0.03)	0.55 (0.13)
Random Forest	0.04 (0.03)	0.70 (0.16)	0.02 (0.01)	0.47 (0.14)	0.03 (0.02)	0.45 (0.08)

PR-AUC: precision-recall area under the curve; ROC-AUC: receiver operating characteristic area under the curve.

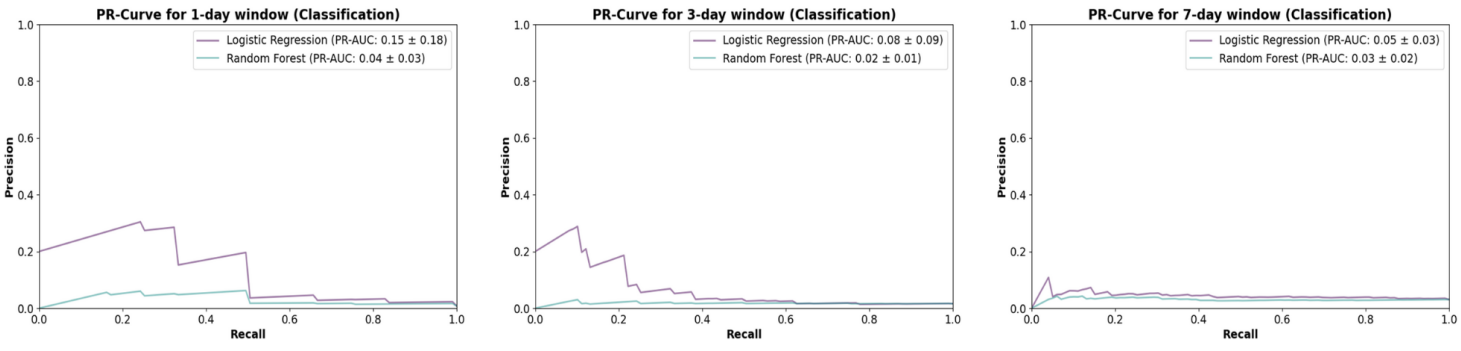


Figure 3.5: Precision-recall curves of the classification models for a 1-day window (left), 3-day window (middle), and 7-day window (right) as the outcome variable.

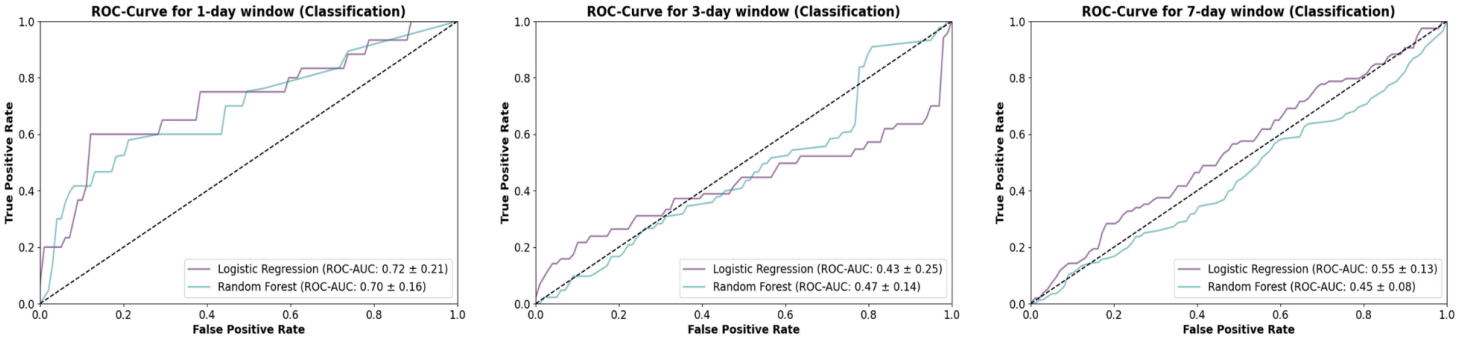


Figure 3.6: Receiver operating characteristic curves of the classification models for a 1-day window (left), 3-day window (middle), and 7-day window (right) as the outcome variable.

3.4. Secondary analysis

3.4.1. Symptom days (ACD-6 ≥ 1.5 and CF symptom score ≥ 7)

Using the individual population sets, the heightened symptom days (indicated by ACD-6 ≥ 1.5 or CF symptom score (SS) ≥ 7) were predicted, including one day prior to the occurrence (1-day window). The number of instances for this prediction was 21.5% (338/1570 days) for ACD-6 ≥ 1.5, and 33.6% (279/831 days) for CF symptom score ≥ 7. The PR-AUC and ROC-AUC for these analyses are shown in Table 3.4, with their corresponding PR-curves and ROC-curves in Appendix A.6.

Table 3.4: Performance evaluation results (PR-AUC and ROC-AUC) for the prediction of symptom days (ACD-6 ≥ 1.5 & CF symptom score ≥ 7) including one day prior to occurrence (1-day window).

	Symptom Days			
	ACD-6 ≥ 1.5		CF symptom score ≥ 7	
	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC
Anomaly Detection				
GMM + PCA	0.25 (0.06)	0.52 (0.01)	0.38 (0.06)	0.53 (0.07)
Isolation Forest	0.26 (0.06)	0.54 (0.03)	0.38 (0.04)	0.51 (0.04)
One-class SVM	0.25 (0.06)	0.52 (0.04)	0.36 (0.04)	0.51 (0.04)
Local Outlier Factor	0.24 (0.06)	0.53 (0.04)	0.36 (0.06)	0.50 (0.07)
Classification				
Logistic Regression	0.28 (0.09)	0.55 (0.05)	0.37 (0.08)	0.50 (0.05)
Random Forest	0.27 (0.11)	0.55 (0.06)	0.32 (0.08)	0.46 (0.12)

ACD-6: six-question Asthma Control Diary; CF: cystic fibrosis; PR-AUC: precision-recall area under the curve; ROC-AUC: receiver operating characteristic area under the curve; GMM: gaussian mixture model; SVM: support vector machine.

The performances in Table 3.4 and Appendix A.6 show that no model significantly performs better than the other. However, the standard deviations of the random forest models were higher

compared to those of the anomaly detection models. Additionally, while the ROC-AUC values were similar for predicting both questionnaire scores, a higher PR-AUC was observed for the CF symptom score, in comparison to the ACD-6 scores.

3.4.2. Symptom days (variable ACD-6 and CF symptom score)

As questionnaire scores are subjective, the analysis was extended to different threshold values that define heightened symptom days, to assess how varying these thresholds impacts the model's performance. The models were evaluated using ACD-6 thresholds of [0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7] and CF symptom score thresholds of [4, 5, 6, 7, 8], while maintaining a 1-day prediction window. The performance results for these analyses are visualised in Figures 3.7 and 3.8, with additional information on the number of instances provided in Appendix A.6

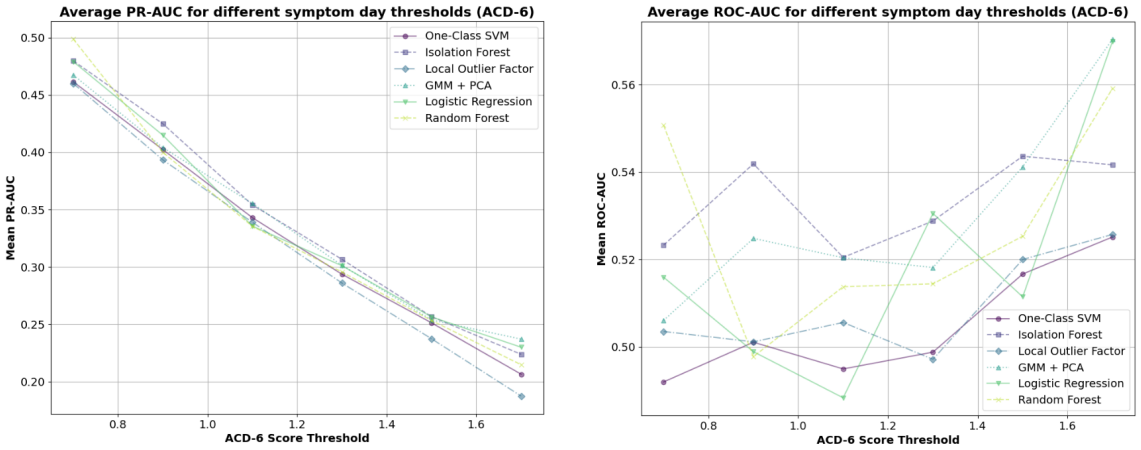


Figure 3.7: PR curves and ROC curves for different minimum thresholds of the heightened symptom days (ACD-6).

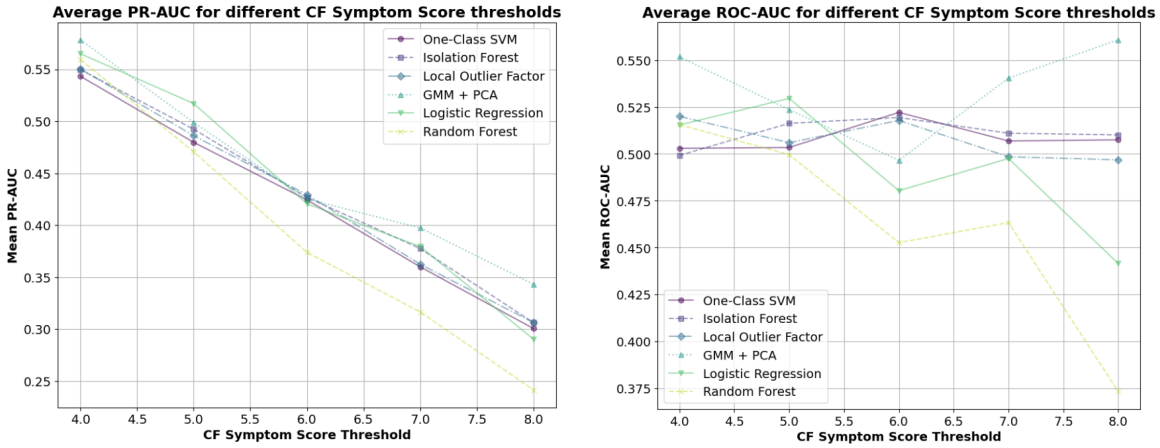


Figure 3.8: PR curves and ROC curves for different minimum thresholds of the heightened symptom days (CF Symptom Score (SS))

Table A.4 in the Appendix shows that lower the threshold was set, the more days were predicted as heightened symptoms days. For the prediction of the ACD-6 scores, lower thresholds led to improved PR-AUC values across all models. However, the ROC-AUC did not show similar improvements. The ROC-AUC generally increased at higher thresholds, as seen with the threshold set at ACD-6 ≥ 1.7. In the case of the CF symptom scores, both the PR-AUC and ROC-AUC improved at lower threshold values for the classification models. However, while the PR-AUC also increased at lower thresholds for the anomaly detection models, the ROC-AUC remained relatively stable at all threshold values.

3.4.3. Additional outcome variables

Lastly, to validate the robustness and generalisability of the methodology, the dataset was investigated to see what additional predictions could be made, even if these outcomes might not have direct clinical relevance. Predictions were made for the following outcome variables: weekend, holiday, and wake-up count ≥ 3. The ROC-AUC results are visualised in Figure 3.9, with further details on the feature importances in Appendix A.6. Only the ROC-AUC scores were shown as there was no class imbalance.

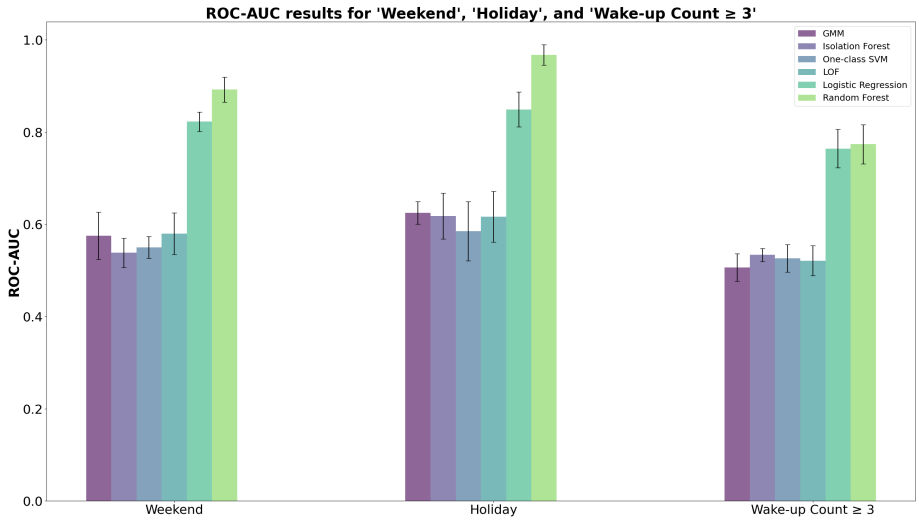


Figure 3.9: Bar plot of the ROC-AUC scores for the prediction of the outcomes 'Weekend', 'Holiday', 'Wake-up count ≥ 3' using the classifiers Gaussian Mixture Model (GMM), Isolation Forest, One-class SVM, Local Outlier Factor (LOF), Logistic Regression, and Random Forest.

Figure 3.9 shows a discrepancy in performance between the anomaly detection models and the classification models. The highest performances were achieved with the random forest model, with ROC-AUC values ranging from 0.77 to 0.97 and small standard deviations for all three prediction outcomes. In contrast, the anomaly detection models performed similarly across the outcome variables, with ROC-AUCs around 0.60. The feature importances revealed that while *steps08* was the most important predictor for the weekend and holiday outcomes, the prediction of the wake-up count heavily relied on the feature *awakeDuration*.

4

Discussion

In this retrospective study, the primary thesis aim was to predict short-term pulmonary exacerbations using machine learning algorithms based on remote patient monitoring data, for the potential to provide real-time warnings to enable timely intervention and prevention. Both anomaly detection algorithms (Gaussian Mixture Model, Random Forest, One-class SVM, and Local Outlier Factor) and classification algorithms (Logistic Regression and Random Forest) were fitted to the data to create prediction models. Different exacerbation time windows were considered for the outcome variables (1-day, 3-day, and 7-day window) as the short-term prediction period. The results showed that the mean performances of the PR-AUC were less than 0.20, and the ROC-AUC was between 0.43-0.72, indicating that the current performance of these models is not yet sufficient for clinical implementation, likely due to the limitations of the available data.
For the secondary analysis, additional outcomes such as the heightened symptom days (based on the ACD-6 and CF symptom scores) were predicted with the developed models. Similarly to the primary analysis, performance metrics scored low PR-AUC values of 0.19-0.58 and for the ROC-AUC values between 0.37-0.58, indicating that the model is no better than random guessing (ROC-AUC = 0.50). This analysis confirms that the performance of the models is not yet sufficient to predict monitoring days related to the symptom disease. Despite this, the methodology of this study was validated by predicting non-clinical variables of the data set. Outcome variables such as weekend or holiday were able to be predicted with good performances (ROC-AUC > 0.80).

4.1. Interpretation of results

The results of the baseline performances showed that in a controlled environment with synthetic data, logistic regression may be more suitable for distinguishing anomalies. A good performance of PR-AUC > 0.80 was reached at a minimum mean anomaly shift of 0.5, indicating that the outliers should differ by at least 0.5 units from the inliers to achieve better results. How-

ever, as this baseline performance used a synthetic dataset, the anomaly shift should be seen as an idealised benchmark.

For the prediction of exacerbations in the primary analysis, the performance metrics showed continuous low/moderate performances. Even though the ROC-AUC showed moderate performances (<0.75), the standard deviation (±0.20) suggested variability in model performances and that the model is likely overfitting on the data. Expanding the window frame of the exacerbation period did not improve the metrics. On the contrary, the highest ROC-AUC was found using the one-day window outcome, for logistic regression (0.72±0.21). Furthermore, the combination of moderate/high ROC-AUC and low PR-AUC could likely be attributed to the imbalanced nature of the dataset, in which the model performs better on the majority class, than the minority class. When handling extremely skewed datasets, ROC-AUC can sometimes present an overly optimistic view of the model's performance. Therefore, the model performed poorly in correctly identifying exacerbations up to one day before occurrence and it is most likely an overfit on the data.

In the secondary analysis, the number of heightened symptom days was predicted prior to one day before occurrence. These results showed improved performances compared to the prediction of exacerbations, but not yet sufficient for the clinical setting. The metrics showed performances with PR-AUC values between 0.24-0.28 and ROC-AUC values between 0.52-0.55 for predicting ACD-6 ≥ 1.5, and PR-AUCs of 0.32-0.38 and ROC-AUCs of 0.46-0.53 for CF symptom score ≥ 7. However, important to note was that the PR-AUCs were close to the baseline PR curve height, which is equal to the proportion of positive examples in the dataset (25.5% for ACD-6 and 37.5% for CF symptom scores). This suggests that the models' predictions are not much better than simply guessing based on the prevalence of the outcome. Furthermore, the ROC-AUCs for both predictions were also near 0.50, further indicating that the models struggled to differentiate between true positives and false positives effectively.
This was confirmed by the results of the variable thresholds depicting heightened symptom days. The average PR-AUC only increased due to the increase in positive instances, as the PR-AUC remained equal to the baseline proportion of instances. Additionaly, the ROC-AUC remained around 0.50, indicating that the model was not able to distinguish between the positive and negative classes. Therefore, heightened symptom days based on minimal ACD-6 and CF symptom scores might be too subjective to be reliably predicted using home monitoring data.

Lastly, additional outcomes such as weekend, holiday, and wake-up count ≥ 3 were predicted. The classification models showed clear superiority over the anomaly detection models. Among the classification models, the random forest demonstrated the highest performance across all three outcomes (ROC-AUC > 0.75), with good generalisation as indicated by the high ROC-AUC scores and small standard deviations. These results indicate that, although the models were not effective in predicting exacerbation and heightened symptom days, they performed better than random chance in predicting non-clinically relevant outcomes. This suggests that the methodology is robust and capable of identifying meaningful patterns in the data.

4.2. Comparative work

A recent study from Sutcliffe et al., [32] on predicting exacerbations using home monitoring data in CF patients, showcased that a logistic regression model can detect impending events 10 days earlier than clinical practice, with an 83.6% success rate at a false positive rate of 18.6%. Out of 15966 active study days, 111 exacerbation events (0.7%) were observed. The author demonstrated that symptom features such as wellness, O2 saturation, and pulse rate, give considerable prediction value, suggesting that home monitoring might not have to involve spirometry, which is effort-dependent, time-consuming, and can cause discomfort to patients.
The success rate mentioned by the author was based on a custom performance metric. Performance evaluation using the standard metrics resulted in a PR-AUC of 0.28, and a ROC-AUC of 0.74%. These findings show that even with larger datasets, correctly identifying the minority class remains a challenge.

In another recent study, Hond et al., [56] showed the superiority of using logistic regression over other ML classifiers, such as XGBoost, for the prediction of asthma exacerbations using remote patient monitoring data. Logistic regression achieved better performances with a ROC-AUC of 0.88, and better sensitivity and specificity. The rate of incidence was 154 exacerbations (0.2 % of total daily measurements) for the development cohort and 94 exacerbations (also 0.2 % of total daily measurements) for the validation cohort. According to the author, the logistic regression classifier had a substantial number of false positive predictions at high levels of sensitivity, which could be linked to the low incidence rate.
Similar to the results of this thesis, the precision would quickly drop at higher recalls, most likely due to severe class imbalance and large variety per patient. Furthermore, Hond performed a sensitivity analysis on expanding the exacerbation outcome window from two to four and eight days, which showed no noticeable performance differences. Although the results in this thesis showed slightly better performances for the 1-day window outcome, the high standard deviation suggested considerable variability and poor generalisation.

4.3. Study limitations

A major limitation of this thesis was the small amount of exacerbations recorded in the monitoring period. Each patient was monitored for 28 days, but out of the 90 included patients, only 11 exacerbation events were observed, one of which was excluded due to insufficient wear time of the smartwatch. Such a small positive class heavily influences the outcome of the algorithms, and generally, large amounts of data are needed for training to generalise well for unseen data in a ML model [57]. No upsampling or downsampling was performed as the number of positive instances was too small, and there is no proven strategy for employing such methods without distorting information and/or introducing bias [58].
Another limitation to consider was that this study was a *post hoc* analysis of remote patient monitoring data, collected for a study aimed at the clinical validation of smartwatch biomarkers, and not originally intended for predicting pulmonary exacerbations. Therefore, the quality required for a reliable ML prediction model was lacking, which contributed to the performance

results. Additionally, the original dataset with continuous measurements was not available. Instead, the acquired dataset consisted of the averaged hourly features for the heart rate and number of steps taken. This limited the application of pre-processing steps, such as filtering of the heart rate data to compensate for sensor inaccuracies.
Moreover, the reliability of the spirometry features was a limiting factor, due to the high proportion of missing or low-quality data. Approximately 39.3% of the spirometry data was either missing or deemed inadequate for analysis. Although KNN-imputation was performed, the reliance on imputed data may have introduced additional uncertainty into the models.
The PR-AUC and ROC-AUC values were close to random guessing, which implied that the models were not able to capture any meaningful patterns from the features. In particular, high feature importance for spirometry features might be misleading. The models may have overfitted on noise or artifacts in the data rather than identifying genuinely predictive patterns, especially given the poor overall model performance. Additionally, feature importance was only employed for the random forest model. To improve the explainability of the other models, additional methods such as the SHapley Additive exPlanations (SHAP) tool may be required [59].
The final limitation was the definition of an exacerbation. As mentioned before, exacerbations are patient-specific and there is no clear consensus on what clinical criteria an exacerbation constitutes. This study had two definitions of exacerbations: worsening of symptoms requiring systemic corticosteroids (asthma patients), and the need for additional antibiotic treatment due to change in symptoms or decrease in pulmonary function (CF patients). Exacerbation time windows were formed as outcome variables, to encompass the change in symptoms before the occurrence of an event. However, the additional need for therapy is subjective, and patient data is heterogeneous. A fixed time window might therefore not capture all the variability [32].

4.4. Recommendations

Given the outcomes of this study, many steps still have to be taken before wearables and AI can be implemented in the clinical setting for predicting pulmonary exacerbations in paediatric care. There is a need for large high-quality longitudinal studies to evaluate the feasibility of passive monitoring and exacerbation prediction. Importantly, for more accurate prediction, an increased number of observed exacerbation events is needed, which can be achieved by extending the monitoring period and/or including more patients. These studies should incorporate multiparameter monitoring strategies, with devices that cause limited patient burden, in particular for children. Furthermore, additional monitoring features could be explored, such as the respiratory rate, O2 saturation, nocturnal cough (physiological), inhaler usage, and patient-reported outcomes (clinical) [60–62]. Building on this framework, features derived from the raw data can be feature-engineered into higher-level behavioral markers (e.g., stress, fatigue, sleep disruption) [63], and more advanced time series analysis can be performed to identify pattern trends and dependencies. Furthermore, the results suggest that anomaly detection models, particularly GMM, may lack the stability and precision required for accurately predicting exacerbations. In datasets characterised by complex features, missing data, variable patient populations, biased

datasets, and heterogeneous outcomes, standard machine learning approaches may struggle to generalise effectively [64, 65]. In such cases, deep learning methods, like Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs), which are adept at handling temporal dependencies, may offer a more suitable approach for predicting exacerbations. However, these methods require enormous large datasets to be effective.
Moreover, in contrast to using time windows to predict the exacerbation events as a category (yes or no), it might be more suitable in the future to predict the probabilities of events, indicating the likelihood of something happening, as explored by two previous studies [66, 67]. This approach was not yet employed in the current study, due to the initial need for a clear, binary, and easily interpretable assessment of the model's performance. Once the models achieve reliable performance with binary predictions, probabilistic outputs may be considered to provide further benefits. This more complex algorithm may provide the prediction as a risk score, which might be more intuitive for both the patient and healthcare professionals to better understand the status of the pulmonary condition.

Lastly, in the future realm of performing real-time predictions of exacerbation events, an ideal architecture system should include several key components. A wearable device, including a smartphone, should be used for continuous data collection. This information should be securely sent to a central server, in which data processing can be conducted, and the model algorithm is updated with new patient information. Initially, this model may only rely on population-based algorithms. However, as the system continues to collect patient-specific data, it should enable the constant retraining and improvement of the underlying algorithms. Over time, this process will allow the transition from population-based models to more personalised predictive algorithms uniquely tailored to the individual patient, resulting in more accurate and personalised exacerbation predictions. Ideally, this system would also be integrated with the hospital's electronic health record, ensuring that any new information documented in the records is incorporated into the model, while measurements and predicted events are made accessible to both patients and healthcare professionals.

4.5. Conclusion

This thesis highlights the need for high-quality data for utilising machine learning in the prediction of pulmonary exacerbations. Currently available patient monitoring data including physiological data, lung function parameters, environmental data, and patient-reported outcomes do not suffice to predict pulmonary exacerbations within 7 days or heightened symptom days. Clinical application may be challenging due to the low incidence rate of exacerbations. The use of machine learning and wearable technology holds significant potential for improving the management of pulmonary exacerbations in pediatric patients, however many improvements in data collection, outcome definition, model development, and evaluation are needed before a well-generalised prediction model can be formed.

Bibliography

[1] Vanhommerig JW, Poos MJJC, Gommer AM, Hendriks C, Wijga AH, Hilderink HBM, et al.. Astma | Volksgezondheid en Zorg — vzinfo.nl; 2022. [Accessed 06-05-2024]. `https://www.vzinfo.nl/astma`.

[2] Devonshire AL, Kumar R. Pediatric asthma: Principles and treatment. In: Allergy & Asthma Proceedings. vol. 40; 2019. .

[3] Papi A, Brightling C, Pedersen SE, Reddel HK. Pathogenesis of asthma. Lancet. 2018;391:783-800.

[4] Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD. Genetics of Asthma and Hay Fever in Australian Twins1-3. Am rev respir Dis. 1990;142:1351-8.

[5] Noordhoek JJ, Zomer DD, Okhuijsen R, Altenburg J, Bannier MAGE, Bakker M, et al.. CF Registratie - NCFS — ncfs.nl; 2022. [Accessed 27-05-2024]. `https://ncfs.nl/onderzoek-naar-taaislijmziekte/cf-registratie/`.

[6] Büscher R, Grasemann H. Disease modifying genes in cystic fibrosis: therapeutic option or one-way road? Naunyn-Schmiedeberg's archives of pharmacology. 2006;374:65-77.

[7] Elborn JS, Bell SC, Madge SL, Burgel PR, Castellani C, Conway S, et al. Report of the European Respiratory Society/European Cystic Fibrosis Society task force on the care of adults with cystic fibrosis. European Respiratory Journal. 2016;47(2):420-8.

[8] Veit G, Avramescu RG, Chiang AN, Houck SA, Cai Z, Peters KW, et al. From CFTR biology toward combinatorial pharmacotherapy: expanded classification of cystic fibrosis mutations. Molecular biology of the cell. 2016;27(3):424-33.

[9] Bilton D, Canny G, Conway S, Dumcius S, Hjelte L, Proesmans M, et al. Pulmonary exacerbation: towards a definition for use in clinical trials. Report from the EuroCareCF Working Group on outcome parameters in clinical trials. Journal of Cystic Fibrosis. 2011;10:S79-81.

[10] Manti S, Licari A, Leonardi S, Marseglia GL. Management of asthma exacerbations in the paediatric population: a systematic review. European Respiratory Review. 2021;30(161).

[11] Papadopoulos NG, Christodoulou I, Rohde G, Agache I, Almqvist C, Bruno A, et al. Viruses and bacteria in acute asthma exacerbations–A GA2LEN-DARE* systematic review. Allergy. 2011;66(4):458-68.

[12] Etzel RA. How environmental exposures influence the development and exacerbation of asthma. Pediatrics. 2003;112(Supplement_1):233-9.

[13] Sarikloglou E, Fouzas S, Paraskakis E. Prediction of Asthma Exacerbations in Children. Journal of Personalized Medicine. 2023;14(1):20.

[14] Jang J, Chan KCG, Huang H, Sullivan SD. Trends in cost and outcomes among adult and pediatric patients with asthma: 2000–2009. Annals of Allergy, Asthma & Immunology. 2013;111(6):516-22.

[15] Ferrante G, La Grutta S. The burden of pediatric asthma. Frontiers in pediatrics. 2018;6:186.

[16] Waters V, Ratjen F. Pulmonary exacerbations in children with cystic fibrosis. Annals of the American Thoracic Society. 2015;12(Supplement 2):S200-6.

[17] Fuhlbrigge A, Peden D, Apter AJ, Boushey HA, Camargo Jr CA, Gern J, et al. Asthma outcomes: exacerbations. Journal of Allergy and Clinical Immunology. 2012;129(3):S34-48.

[18] Goss CH, Burns JL. Exacerbations in cystic fibrosis· 1: epidemiology and pathogenesis. Thorax. 2007;62(4):360-7.

[19] Global Strategy for Asthma Management and Prevention (2024 update). Global Initiative for Asthma (GINA); 2024. `https://ginasthma.org/reports/`.

[20] Passos S, Maziero FF, Antoniassi DQ, de Souza LT, Felix AF, Dotta E, et al. ACUTE RESPIRATORY DISEASES IN BRAZILIAN CHILDREN: ARE CAREGIVERS ABLE TO DETECT EARLY WARNING SIGNS? Revista Paulista de Pediatria. 2018;36:3 9.

[21] Zorgstandaard Astma Kinderen & Jongeren. Long Alliantie Nederland; 2012.

[22] Sarasohn-Kahn J. The connected patient: charting the vital signs of remote health monitoring. California HealthCare Foundation; 2011.

[23] Loeb AE, Rao SS, Ficke JR, Morris CD, Riley III LH, Levin AS. Departmental experience and lessons learned with accelerated introduction of telemedicine during the COVID-19 crisis. JAAOS-Journal of the American Academy of Orthopaedic Surgeons. 2020;28(11):e469-76.

[24] Davies B, Kenia P, Nagakumar P, Gupta A. Paediatric and adolescent asthma: a narrative review of telemedicine and emerging technologies for the post-COVID-19 era. Clinical & Experimental Allergy. 2021;51(3):393-401.

[25] Condry MW, Quan XI. Remote Patient Monitoring Technologies and Markets. IEEE Engineering Management Review. 2023.

[26] Mohammadzadeh N, Gholamzadeh M, Saeedi S, Rezayi S. The application of wearable smart sensors for monitoring the vital signs of patients in epidemics: a systematic literature review. Journal of ambient intelligence and humanized computing. 2023:1-15.

[27] Trifan A, Oliveira M, Oliveira JL, et al. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. JMIR mHealth and uHealth. 2019;7(8):e12649.

[28] Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR mHealth and uHealth. 2020;8(9):e18694.

[29] van der Kamp MR, Hengeveld VS, Brusse-Keizer MG, Thio BJ, Tabak M. eHealth Technologies for Monitoring Pediatric Asthma at Home: Scoping Review. Journal of Medical Internet Research. 2023;25:e45896.

[30] Hosseini A, Buonocore CM, Hashemzadeh S, Hojaiji H, Kalantarian H, Sideris C, et al. Feasibility of a secure wireless sensing smartwatch application for the self-management of pediatric asthma. Sensors. 2017;17(8):1780.

[31] Tsang KC, Pinnock H, Wilson AM, Salvi D, Shah SA. Home monitoring with connected mobile devices for asthma attack prediction with machine learning. Scientific Data. 2023;10(1):370.

[32] Sutcliffe D, Ukor E, Ryan J, Allen J, Brown K, Bell N, et al. Machine learning predicts acute pulmonary exacerbations in cystic fibrosis. 2021.

[33] Shaik T, Tao X, Higgins N, Li L, Gururajan R, Zhou X, et al. Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2023;13(2):e1485.

[34] Liu VWF. Machine learning for prediction of pulmonary exacerbations using medical devices in remote patient monitoring: a literature review; 2024.

[35] Kruizinga MD, Essers E, Stuurman FE, Yavuz Y, de Kam ML, Zhuparris A, et al. Clinical validation of digital biomarkers for paediatric patients with asthma and cystic fibrosis: potential for clinical trials and clinical care. European Respiratory Journal. 2022;59(6).

[36] Bhatt JM. Treatment of pulmonary exacerbations in cystic fibrosis. European Respiratory Review. 2013;22(129):205-16.

[37] Graham BL, Steenbruggen I, Miller MR, Barjaktarevic IZ, Cooper BG, Hall GL, et al. Standardization of spirometry 2019 update. An official American thoracic society and European respiratory society technical statement. American journal of respiratory and critical care medicine. 2019;200(8):e70-88.

[38] scikit-learn developers. LabelEncoder; 2024. Available from: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html`.

[39] scikit-learn developers. StandardScaler; 2024. Available from: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`.

[40] scikit-learn developers. KNNImputer; 2024. Available from: `https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html`.

[41] Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, Maconochie I, et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. The Lancet. 2011;377(9770):1011-8.

[42] van der Kamp MR, Klaver EC, Thio BJ, Driessen JM, de Jongh FH, Tabak M, et al. WEARCON: wearable home monitoring in children with asthma reveals a strong association with hospital based assessment of asthma control. BMC medical informatics and decision making. 2020;20:1-12.

[43] Brownlee J. Basic Feature Engineering With Time Series Data in Python; 2021. Available from: `https://machinelearningmastery.com/basic-feature-engineering-time-series-data-python/`.

[44] Reynolds DA, et al. Gaussian mixture models. Encyclopedia of biometrics. 2009;741(659-663).

[45] Wan H, Wang H, Scotney B, Liu J. A novel gaussian mixture model for classification. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE; 2019. p. 3298-303.

[46] Liu FT, Ting KM, Zhou ZH. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. IEEE; 2008. p. 413-22.

[47] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.. sklearn.ensemble.IsolationForest; 2011. Accessed: 2024-07-06. Available from: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html`.

[48] Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural computation. 2001;13(7):1443-71.

[49] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data; 2000. p. 93-104.

[50] Schober P, Vetter TR. Logistic regression in medical research. Anesthesia & Analgesia. 2021;132(2):365-6.

[51] Stoltzfus JC. Logistic regression: a brief primer. Academic emergency medicine. 2011;18(10):1099-104.

[52] Breiman L. Random forests. Machine learning. 2001;45:5-32.

[53] Lee Hj, Cho S. The novelty detection approach for different degrees of class imbalance. In: International conference on neural information processing. Springer; 2006. p. 21-30.

[54] Novelty Detection Definition | DeepAI;. Accessed: 2024-06-22. https://deepai.org/machine-learning-glossary-and-terms/novelty-detection.

[55] Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010;2(4):433-59.

[56] de Hond AA, Kant IM, Honkoop PJ, Smith AD, Steyerberg EW, Sont JK. Machine learning did not beat logistic regression in time series prediction for severe asthma exacerbations. Scientific reports. 2022;12(1):20363.

[57] Tomalin M, Byrne B, Concannon S, Saunders D, Ullmann S. The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. Ethics and Information Technology. 2021;23:419 433.

[58] Sabha SU, Assad A, Din NMU, Bhat M. Comparative Analysis of Oversampling Techniques on Small and Imbalanced Datasets Using Deep Learning. 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP). 2023:1-5.

[59] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

[60] Tsang KC, Pinnock H, Wilson AM, Shah SA. Application of machine learning to support self-management of asthma with mHealth. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE; 2020. p. 5673-7.

[61] Cooper CB, Sirichana W, Arnold MT, Neufeld EV, Taylor M, Wang X, et al. Remote patient monitoring for the detection of COPD exacerbations. International Journal of Chronic Obstructive Pulmonary Disease. 2020:2005-13.

[62] Tinschert P, Rassouli F, Barata F, Steurer-Stey C, Fleisch E, Puhan MA, et al. Nocturnal cough and sleep quality to assess asthma control and predict attacks. Journal of asthma and allergy. 2020:669-78.

[63] Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annual review of clinical psychology. 2017;13(1):23-47.

[64] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence. 2013;35(8):1798-828.

[65] Stewart C. Understanding disease through remote monitoring technology. King's College London; 2024.

[66] Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmy L, et al. Asthma exacerbation prediction and interpretation based on time-sensitive attentive neural network: A retrospective cohort study. medRxiv. 2019:19012161.

[67] Do Q, Tran S, Doig A. Reinforcement learning framework to identify cause of diseases-predicting asthma attack case. In: 2019 IEEE International Conference on Big Data (Big Data). IEEE; 2019. p. 4829-38.

A

Appendix

A.1. Feature Engineering

Three features related to the heart rate were constructed. The first feature, nocturnal heart rate reserve, was calculated to capture the difference between the maximum heart rate during sleep and the average heart rate during the night. This feature was based on the feature importance of the heart rate reserve, as described by Hosseini et al. [30]. The nocturnal heart rate reserve was calculated as follows:

$$\text{Nocturnal Heart Rate Reserve} = \text{HR}_{\text{Max Sleep}} - \text{HR}_{\text{Avg}}$$

The second feature was composed to analyse deviations from age-specific normal paediatric heart rates as reported by Fleming et al. [41]. This feature was calculated for the daily average heart rates, the hourly heart rates, the 95th percentile, and the 5th percentile heart rates. The differences were calculated as follows:

$$\text{Daily Heart Rate Difference} = \text{HR}_{\text{Measured}} - \text{HR}_{\text{Normal Adjusted}}$$

Lastly, a clinical feature capturing the relationship between physical activity and heart rate was constructed. According to Mathienne et al. [42], the recovery time for the heart rate after exercise was higher in patients with uncontrolled asthma. To capture the relationship between the heart rate and activity, the following ratio was calculated:

$$\text{Ratio Heart Rate and Steps} = \frac{\text{Average Heart Rate During Hour X}}{\text{Total Number of Steps in Hour X}}$$

This feature was based on the ratio between the highest number of steps taken in an hour and the corresponding average heart rate at that time. Both the heart rate and number of steps taken were normalised before calculation, to ensure consistency and comparability across patients.

A.2. Hyperparameter settings

The following settings in Table A.1 were considered for hyperparameter tuning during the inner loop of the nCV.

Table A.1: Hyperparameter settings for the anomaly detection and classification models

Model	**Hyperparameter**	**Values**
Gaussian Mixture Model	`n_gaussian`	[1, 2, 3, 4]
	`pca_components`	[5, 10, 25, 50, 75, 100]
One-Class SVM	`nu`	[0.01, 0.05, 0.1]
	`kernel`	[rbf]
Isolation Forest	`n_estimators`	[100, 200]
	`max_samples`	[auto, 0.5, 0.75]
Local Outlier Factor	`n_neighbors`	[10, 20, 35]
Logistic Regression	`pca_components`	[5, 10, 25, 50, 75, 100]
Random Forest	`n_estimators`	[100, 200]
	`max_depth`	[None, 5]
	`class_weight`	['balanced']

A.3. Evaluation Metrics

The evaluation metrics used to evaluate the model's performance are further explained as follows:

Precision-Recall Curve (PR-Curve)

The Precision-Recall Curve (PR-Curve) is a graphical representation of a model's performance, plotting precision (y-axis) against recall (x-axis). Precision and recall are further explained in the next section. In this plot, precision and recall are calculated for different threshold values, showing the trade-off between precision and recall as the threshold changes. Thresholds are necessary for mapping data samples to one class or the other. The algorithms of the models use the thresholds to interpret the mapping of the labels. The default threshold is 0.50, in which values that are less than 0.50 are assigned to class 0, and values larger than or equal to 0.50 are assigned to class 1. The left side of these curves indicates a more 'confident' threshold, with a higher threshold (e.g., threshold = 0.80) corresponding to lower recall but higher precision. The right side represents 'less strict' scenarios, where the thresholds are lower (e.g., threshold = 0.20), with higher recall but lower precision. The area under the PR Curve (PR-AUC) is a single value which summarises the classifier's performance over all threshold values.

Receiver Operating Characteristic Curve (ROC Curve)

The Receiver Operating Characteristic (ROC) Curve plots the true positive rate (sensitivity or recall) against the false positive rate (1 - specificity) at all threshold values, which provides a comprehensive view of this trade-off of these rates. The area under the ROC Curve (ROC-AUC) measures the classifier's ability to distinguish between the classes and is used as a summary of the ROC curve. An ROC-AUC of 0.50 indicates that the model is not able to distinguish between the classes, corresponding to a random classifier.

Precision

Precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision, also known as the positive predictive value, measures the accuracy of all the positive predictions made by the model.

Recall (Sensitivity)

Recall, also known as sensitivity, is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall measures whether all positive instances in the dataset can be correctly identified by the model.

Specificity

Specificity is defined as:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Specificity measures how well all negative instances in the dataset can be identified.

F1-Score

The F1-Score is the harmonic mean between precision and recall, defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-Score is a single value metric that balances both precision and recall, and it is commonly used in cases where the class distribution is imbalanced.

A.4. Primary analysis: Anomaly detection

The extended results of the performance metrics of the anomaly detection models are presented in Table A.2.

Table A.2: Performance metrics (mean and standard deviation) for different anomaly detection models across different time windows.

Anomaly Detection			
	1 day window	**3 day window**	**7 day window**
One-class SVM			
Precision	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Recall	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
F1 Score	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
AUROC	0.649 (0.247)	0.619 (0.170)	0.513 (0.125)
PR AUC	0.042 (0.027)	0.038 (0.019)	0.049 (0.025)
Specificity	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Isolation Forest			
Precision	0.033 (0.075)	0.000 (0.000)	0.056 (0.053)
Recall	0.050 (0.112)	0.000 (0.000)	0.052 (0.053)
F1 Score	0.040 (0.089)	0.000 (0.000)	0.054 (0.053)
AUROC	0.620 (0.247)	0.599 (0.199)	0.556 (0.178)
PR AUC	0.053 (0.037)	0.037 (0.022)	0.055 (0.024)
Specificity	0.987 (0.008)	0.992 (0.004)	0.977 (0.008)
Local Outlier Factor			
Precision	0.008 (0.001)	0.016 (0.003)	0.031 (0.003)
Recall	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
F1 Score	0.016 (0.002)	0.031 (0.006)	0.060 (0.006)
AUROC	0.666 (0.200)	0.548 (0.061)	0.450 (0.101)
PR AUC	0.027 (0.013)	0.025 (0.009)	0.036 (0.014)
Specificity	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
GMM + PCA			
Precision	0.008 (0.001)	0.016 (0.003)	0.031 (0.003)
Recall	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
F1 Score	0.016 (0.002)	0.031 (0.006)	0.060 (0.006)
AUROC	0.668 (0.197)	0.599 (0.172)	0.576 (0.166)
PR AUC	0.033 (0.016)	0.035 (0.016)	0.073 (0.043)
Specificity	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

The inner loops of the nested cross-validation were used for hyperparameter tuning. The results for the hyperparameters on each outer fold, for the model assessment on the 7-day window

outcome variable, are as follows:

One-Class SVM

Fold 1: `nu`: 0.10, `kernel`: rbf
Fold 2: `nu`: 0.10, `kernel`: rbf
Fold 3: `nu`: 0.10, `kernel`: rbf
Fold 4: `nu`: 0.10, `kernel`: rbf
Fold 5: `nu`: 0.10, `kernel`: rbf

Isolation Forest

Fold 1: `n_estimators`: 100, `max_samples`: auto
Fold 2: `n_estimators`: 200, `max_samples`: auto
Fold 3: `n_estimators`: 200, `max_samples`: auto
Fold 4: `n_estimators`: 200, `max_samples`: auto
Fold 5: `n_estimators`: 200, `max_samples`: auto

Local Outlier Factor (LOF)

Fold 1: `n_neighbors`: 35, `algorithm`: auto
Fold 2: `n_neighbors`: 20, `algorithm`: auto
Fold 3: `n_neighbors`: 35, `algorithm`: auto
Fold 4: `n_neighbors`: 35, `algorithm`: auto
Fold 5: `n_neighbors`: 35, `algorithm`: auto

Gaussian Mixture Model (GMM)

Fold 1: `pca_components`: 5, `n_gaussian`: 3
Fold 2: `pca_components`: 5, `n_gaussian`: 2
Fold 3: `pca_components`: 5, `n_gaussian`: 2
Fold 4: `pca_components`: 25, `n_gaussian`: 4
Fold 5: `pca_components`: 5, `n_gaussian`: 3

A.5. Primary analysis: Classification

The extended results of the performance metrics of the classification models are presented in Table A.3.

Table A.3: Performance metrics (mean and standard deviation) for different classification models across different time windows.

Classification Models			
	1 day window	**3 day window**	**7 day window**
Logistic Regression			
Precision	0.200 (0.447)	0.000 (0.000)	0.000 (0.000)
Recall	0.033 (0.075)	0.000 (0.000)	0.000 (0.000)
F1 Score	0.057 (0.128)	0.000 (0.000)	0.000 (0.000)
AUROC	0.723 (0.211)	0.430 (0.253)	0.547 (0.131)
PR AUC	0.155 (0.198)	0.083 (0.097)	0.052 (0.032)
Specificity	0.998 (0.003)	0.999 (0.003)	1.000 (0.000)
Random Forest			
Precision	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Recall	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
F1 Score	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
AUROC	0.695 (0.156)	0.468 (0.142)	0.451 (0.078)
PR AUC	0.044 (0.032)	0.022 (0.011)	0.035 (0.017)
Specificity	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

The inner loops of the nested cross-validation were used for hyperparameter tuning. The results for the hyperparameters on each outer fold, for the model assessment on the 7-day window outcome variable, are as follows:

Logistic Regression

Fold 1: `pca_components`: 5
Fold 2: `pca_components`: 50
Fold 3: `pca_components`: 5
Fold 4: `pca_components`: 5
Fold 5: `pca_components`: 50

Random Forest

Fold 1: `n_estimators`: 100, `max_depth`: None
Fold 2: `n_estimators`: 100, `max_depth`: None
Fold 3: `n_estimators`: 100, `max_depth`: None
Fold 4: `n_estimators`: 100, `max_depth`: None
Fold 5: `n_estimators`: 100, `max_depth`: None

Feature Importance

The mean feature importance for the 5-fold nested cross-validated random forest model for the 1-day window outcome is shown in Figure A.1. In particular, predominantly features related to the heart rate measures were rated as the highest in importance.

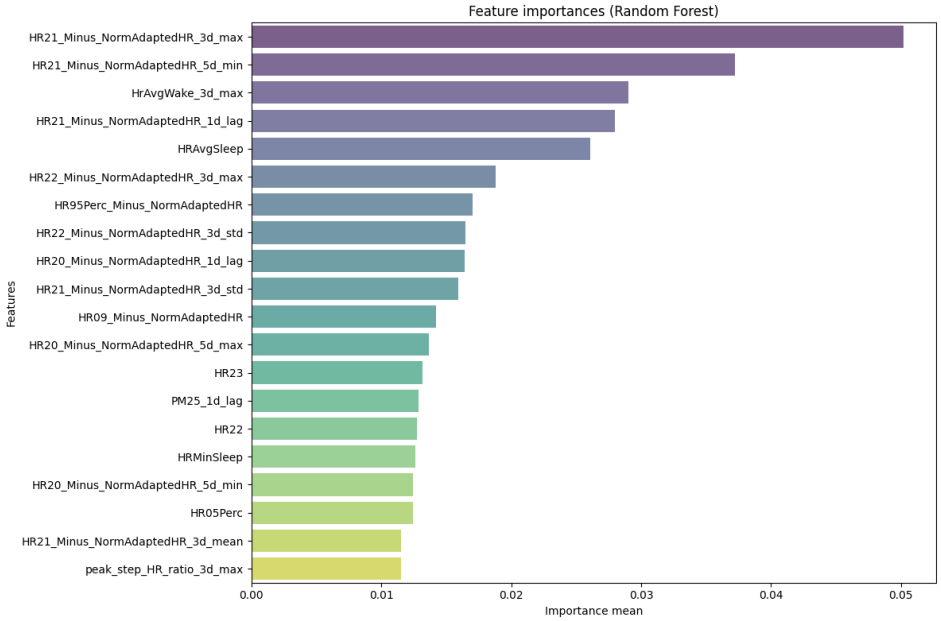


Figure A.1: Mean feature importance scores for the 5-fold nested cross-validated random forest model with 1-day exacerbation window outcome.

A.6. Secondary analysis

Symptom days (ACD-6 ≥ 1.5 and CF symptom score ≥ 7)

Additionally to the PR-AUCs and the ROC-AUCs, the following Figures A.2 and A.3 display the PR-curve and ROC-curve respectively for the asthma and CF symptom days.

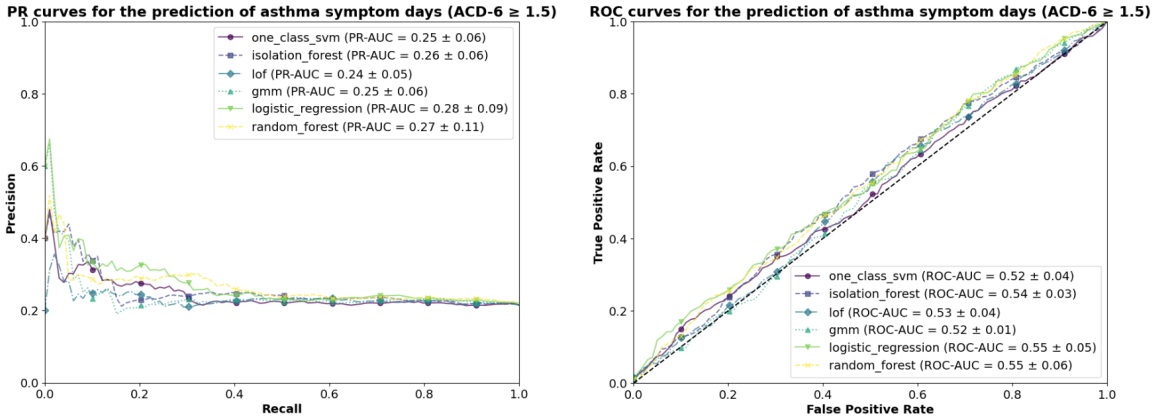


Figure A.2: PR curve (left) and ROC curve (right) of the prediction of an asthma symptom day two days before occurrence, based on six-question Asthma Control Diary Score (ACD-6) ≥ 1.5.

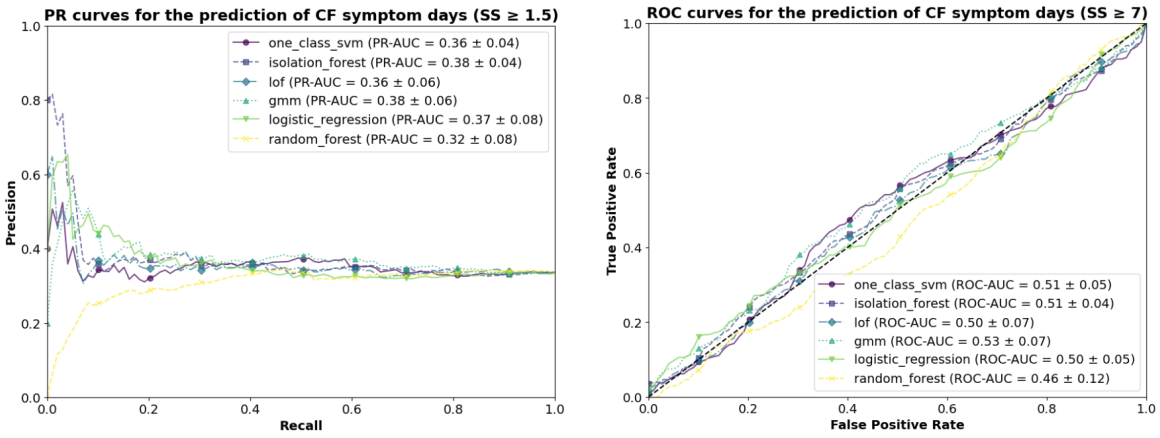


Figure A.3: PR curve (left) and ROC curve (right) of the prediction of a CF symptom day two days before occurrence, based on CF Symptom Score (SS) ≥ 7.

Symptom days (variable ACD-6 and CF symptom score)

The asthma and CF heightened symptom days were predicted based on the mean of the questionnaire scores. The models were evaluated on ACD-6 ≥ [0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7] and CF symptom score ≥ [4, 5, 6, 7, 8]. The number of positive instances of these different threshold are displayed in Table A.4.

Table A.4: Percentage and amount of positive instances for the prediction of symptom day with variable ACD-6 and CF symptom score thresholds.

Positive instances					
	ACD-6			CF Symptom Score	
Threshold			Threshold		
0.7	45,3%	(711/1570)			
0.9	37,9%	(595/1570)	4	51,7%	(430/831)
1.1	32,3%	(506/1570)	5	45,0%	(374/831)
1.3	26,8%	(421/1570)	6	39,0%	(324/831)
1.5	21,5%	(338/1570)	7	33,6%	(279/831)
1.7	15,9%	(248/1570)	8	27,8%	(231/831)

Symptom days (Additional outcome variables)

Additional outcome variables were predicted using the anomaly detection models and classification models. This included the outcomes: Weekend, Holiday, and Wake-up Count ≥ 3. The feature importances for these outcome variables are shown in Figure A.4.

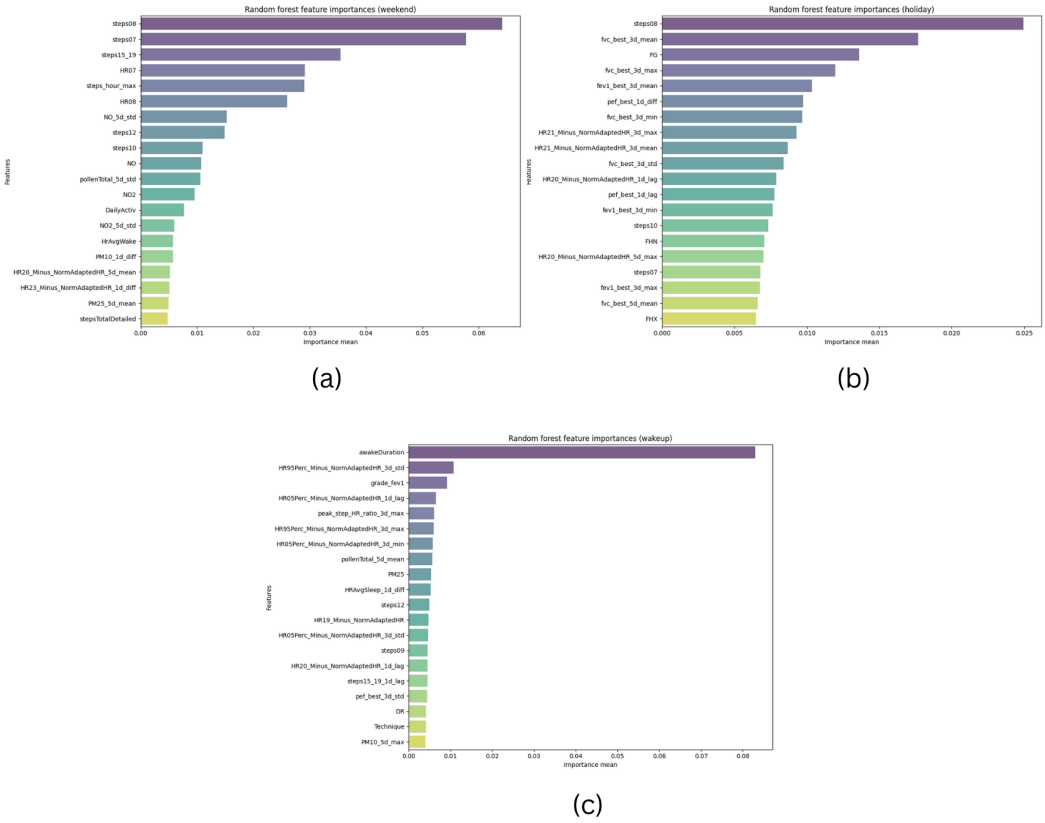


Figure A.4: Random forest feature importances of the outcome variables: a) Weekend, b) Holiday, and c) Wake-up count ≥ 3

For both the weekend and holiday outcomes, the feature *steps08* emerged as the most important predictor. This likely reflected the level of morning activity, as patients are typically more active during this hour due to school, compared to weekends or holidays. In the case of the weekend outcome, activity and heart rate related features were the most important predictors, reflecting changes in physical activity and heart rate patterns typically associated with weekends. For the holiday outcome, however, the features with higher importance included activity, spirometry, and environmental factors, suggesting that these variables play a more significant role in distinguishing holidays from regular days. In the case of the wake-up count ≥ 3 outcome, the feature *awakeDuration* overwhelmingly dominated the feature importance scores. This indicates that the model heavily relied on this feature, which is likely directly related to or used in calculating the wake-up count.