COPD Severity Prediction in Elderly with ML Techniques

Elias Dritsas, Sotiris Alexiou and Konstantinos Moustakas
dritsase@ceid.upatras.gr,salexiou@ceid.upatras.gr,moustakas@ece.upatras.gr
University of Patras, Electrical and Computer Engineering Department
Patra, Rion, Greece

ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a disease characterized by persistent symptoms mainly in the respiratory system and permanent restriction of airflow. It can worsen over time and develop into a serious illness, being one of the leading causes of morbidity and mortality worldwide. In the context of this study, we focus on the early prediction of the COPD patients' severity grades, especially those over 55 years of age. For this purpose, we employ Machine Learning (ML) techniques in order to design appropriate models that will efficiently estimate the severity level based on the most crucial risk factors for disease development. These models will be embedded in the AI Framework of the GATEKEEPER system, which aims to provide personalized risk assessment and interventions to the elderly population.

CCS CONCEPTS

• **Computing methodologies → Machine Learning**.

KEYWORDS

COPD, risk prediction, elderly, healthy life

ACM Reference Format:

Elias Dritsas, Sotiris Alexiou and Konstantinos Moustakas. 2022. COPD Severity Prediction in Elderly with ML Techniques. In *The 15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3529190.3534748

1 INTRODUCTION

Chronic Obstructive Pulmonary Disease is the fourth leading cause of death worldwide. The main cause of the disease is smoking. Less common causes are inhalation of toxic substances or gases (secondhand smoke, air pollution, occupational exposure) and, less frequently, genetic predisposition [1] [2].

In the coming decades, the incidence of COPD is expected to increase due to the constant exposure of humans to harmful agents and the ageing of the population. Significant exposure of the lungs to harmful particles or gases has a detrimental effect on the airways of the lungs and the alveoli of the lungs. Oxygen and carbon dioxide are exchanged between the blood and the respiratory air in the alveoli of the lungs [1].

The disease is characterized by restricted airflow, and patients at an advanced stage have difficulty breathing. The development of the disease is characterized by four stages: mild, moderate, severe and very severe [3].

Shortness of breath is one of the most characteristic symptoms of COPD. It usually has chronic and progressive, gradual deterioration. Cough is also an important symptom. Other symptoms of COPD are tightness in the chest and the presence of sibilation, which is one of the hearing findings during the clinical examination. Other symptoms include fatigue, anorexia nervosa and weight loss. In any case, it is important for the patient to seek medical evaluation for both the symptomatology and the differential diagnosis of other diseases that may present with a similar clinical picture [19].

COPD often coexists with other diseases. Its coexistence with other diseases such as cardiovascular disease, heart failure, arrhythmias, hypertension, lung cancer, diabetes, osteoporosis, depression and gastroesophageal reflux disease can have a negative impact on its course but also on mortality from it [4].

Machine learning is now an important tool in the medical field, as it can help predict various diseases such as diabetes (as classification [10] or regression task for continuous glucose prediction [2], [6]), hypertension [8], cholesterol [9], sleep disorders [14], CVDs [7], stroke [12], COVID-19 [17] etc. In particular, COPD will concern us in the context of this study. For this specific disease, numerous research studies have been conducted with the aid of supervised ML models. In [16], the authors used Random forest to identify which features are most important for case identification. In [11], the authors compared the performance of several machine learning approaches for predicting two clinical outcomes among emergency department (ED) patients with asthma or COPD exacerbation. In [18], a systematic study on developing different types of machine learning models, including both deep and non-deep ones, for predicting the readmission risk of COPD patients is performed. In [20], applying machine learning to clinical and quantitative computed tomography (CT) imaging features would improve mortality prediction in COPD. In [22], the authors compared three of the most common machine learning algorithms (decision tree, naive Bayes and Bayesian network) based on the ROC metric in order to increase COPD prediction accuracy. In [13], an attempt has been made to detect COPD patients and, at the same time, it could distinguish the stages such as the early stage of chronic obstructive pulmonary disease patients (ESCP) and the Advanced stage of COPD patients (ASCP).

In the context of this study, a methodology for designing effective multi-class ML classification models is presented to predict the severity of COPD assuming various risk factors related to lung

[1] https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)
[2] https://goldcopd.org/

PETRAE '22, June 29–July 2, 2022, Corfu, GR

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3534748

function, smoking status, the coexistence of other chronic conditions (such as diabetes, hypertension, diabetes, heart disease) and scoring tests that reflect the quality of life of COPD patients. The ultimate goal of the current research is to integrate the evaluated models into the ML framework of the GATEKEEPER system, which aims to develop personalized risk prediction models to improve the well-being of the elderly.

The rest of this paper is organized as follows. Section 2 gives a brief description of the system in which the current methodology will be integrated. Section 3 analyzes the risk factors of the adopted COPD dataset. In addition, Section 4 presents the elaborated approach for predicting the COPD severity levels. Section 5 lists the methods for data pre-processing and feature classification. The evaluation is performed in Section 6. Finally, Section 7 concludes the work and sets future directions.

2 THE GATEKEEPER SYSTEM

The GATEKEEPER system aims to develop AI-powered services for personalised early risk detection and risk assessment. For this purpose, advanced diagnostic and prognostic AI algorithms and their performances are investigated to be used as part of its interventions. In particular, it focuses on the elderly without technical knowledge in order to enhance their well-being (independence and ability over time). The development of the predictive models is based on the availability of the data from the pilots. The context and specifications of such AI-powered services constitute an important part of the system. The strategy of GATEKEEPER is to provide new intelligent tools and algorithms to outline significant risk factors related to older people's well-being, health and treatment adherence, employing modern practices from data science in order to identify abnormalities in a long list of parameters (risk factors) and determine hidden patterns between (big) data collections.

The AI-powered services include medical-based AI algorithms, health activity monitoring and personalized risk detection and assessment in an AI Reasoning Framework. These modules share an AI strategy for developing the AI and ML components. This strategy consists of four stages:

- Stage 1 specifies the modelling requirements (problem formulation) per use case and the design.
- Stage 2 deals with the development and validation (as part of the model selection) of the respective AI/ML-based solutions.
- Stage 3 aims at verifying both the technical and clinical performance of the AI/ML models through a systematic and thoroughly planned process (verification and validation).
- Stage 4 emphasizes the internal and external validation of the models' output. The system performance monitoring and evaluation in real-world contribute to increasing the transparency of a model's results, where new data may be used for model re-training.

3 DATASET DESCRIPTION

Our research was based on a dataset from Kaggle and focused on participants who are over 50 years old. It is a cohort of patients who had been invited to take part in a rehabilitation program to help manage COPD. The number of participants is 100, and all the attributes (18 as input to ML models and 1 for target class) are described as follows:

- **Age** (years): This feature refers to the age of a person who is over 50 years old. It's numerical data.
- **Gender**: This feature refers person's gender. The number of men is 74 while the number of women is 26. It's nominal data.
- **PackHistory**: This feature refers to a person's pack-years smoking, where pack-years is defined as twenty cigarettes smoked every day for one year. It's numerical data.
- **MWT1** (meters): This feature captures the distance the patient walks in 6 minutes in meters (attempt 1). It's numerical data.
- **MWT2** (meters): This feature captures the distance the patient walks in 6 minutes in meters (attempt 2). It's numerical data.
- **MWT1Best** (meters): This feature captures the distance the patient walks in 6 minutes in meters (best attempt). It's numerical data.
- **FEV1** (mL): This feature represents the amount of air the participant can force from the lungs in one second in litres. It's numerical data.
- **FEV1PRED** (%): This feature, expressed as a percentage change of the predicted value of FEV1, is considered an important prognostic factor in the new GOLD classification. It's numerical data.
- **FVC** (mL): This feature represents the total volume of air that the participant can forcibly exhale in one breath. It's numerical data.
- **FVCPRED** (%): This feature expresses the percentage change of the predicted value of FVC. It's numerical data.
- **CAT**: This feature captures the mean score acquired by the COPD Assessment Test. CAT is a scoring system for COPD patients, which provides a simple method for assessing the impact of COPD on the patient's health. It's numerical data.
- **SGRQ**: This feature captures the quality of life index. The St George's Respiratory Questionnaire (SGRQ) is a validated, commonly used questionnaire for measuring the quality of life in patients with chronic obstructive pulmonary disease (COPD). It's numerical data.
- **Smoking**: This feature refers to whether the participant is smoking or not. The percentage of participants who are smoking is 88%. It's nominal data.
- **Diabetes**: This feature refers to whether the participant has been diagnosed with diabetes or not. The percentage of participants who suffer from diabetes is 21%. It's nominal data.
- **Muscular**: This feature refers to whether the participant is muscular or not. The percentage of participants who are muscular is 20%. It's nominal data.
- **Hypertension**: This feature refers to whether this person is hypertensive or not. The percentage of participants who have hypertension is 24%. It's nominal data.
- **AtrialFib**: This feature refers to whether the participant suffers from atrial fibrillation or not. The percentage of participants who have atrial fibrillation is 22%. It's nominal data.

Features	Min	Max	Mean ± std
Age	53	88	69.64 ± 7.803
MWT1	120	688	365.13 ± 106.787
MWT2	120	699	366.05 ± 111.662
MWTBest	120	699	357.74 ± 110.298
FEV1	0.45	3.18	1.521 ± 0.81
FEV1PRED	3.39	102	53.424 ± 26.052
FVC	1.14	5.23	2.777 ± 1.06
FVCPRED	27	132	79.11 ± 24.436
CAT	3	32	19.18 ± 7.817
SGRQ	8.12	75.56	41.729 ± 17.016
PackHistory	1	103	35.52 ± 23.727

Table 1: Statistical Characteristics

- **IHD**: This feature refers to whether the participant suffers from ischemic heart disease or not. The percentage of participants who suffer from ischemic heart disease is 9%. It's nominal data.
- **COPDSEVERITY**: This feature captures a person's COPD severity grade and has 4 categories (mild, moderate, severe and very severe).

Table 1 presents the statistical characteristics of the numerical features in the dataset. In particular, we observe that the mean age of patients is 69.64 ± 7.803. The mean best distance MWTBest is 357.74±110.298. Also, the mean FEV1 and FEV1PRED are 1.521±0.81 and 53.424 ± 26.052, respectively. Moreover, the mean FVC and FVCPRED are 2.777 ± 1.06 and 79.11 ± 24.436, correspondingly. The CAT score ranges from 3 to 32 (the valid range is 0 – 40), and the mean score is 19.18 ± 7.817. A high score denotes a severe impact of COPD on a patient's life. The SGRQ score ranges from 8.12 to 75.56 (the valid range is 0 – 100, where 0 indicates best health and 100 shows worst health), and the mean score is 41.729 ± 17.016. As for the PackHistory, the minimum value is 1, which can be translated into one packet (i.e., 20 cigarettes) per day for one year or half packet (i.e., ten cigarettes) per day for two years. The mean period of smoking packets per year is 35.52 ± 23.727.

In Figures 1, 2 and 3, we depicted the patients' distribution in the different grades of COPD in terms of the age group they belong to, their gender and smoking status. The moderate and severe classes most frequently occur in the older than 75 years age group. Also, a small percentage of participants in the age groups 50-54 and 55-59 suffer from COPD. COPD disease is more prevalent in men than women and among those who smoke. The highest gap between men and women is shown in the severe class. However, a small percentage of patients suffer from COPD without being smokers.

4 MULTI-CLASS CLASSIFICATION METHODOLOGY

The COPD severity prediction is treated as a multi-class classification task. To solve such problems, the one-against-all and one-against-one methods are used in the relevant literature [21]. We assume K classes, which correspond to the K COPD severity levels. Specifically, in the current problem, there are $K = 4$ classes that capture the disease development in patients with COPD assuming

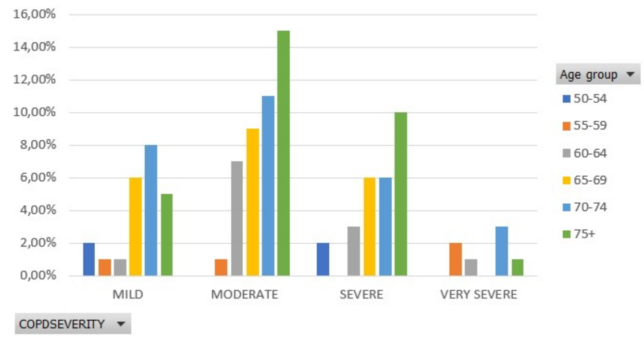

Figure 1: Participants distribution per COPD severity level and age group

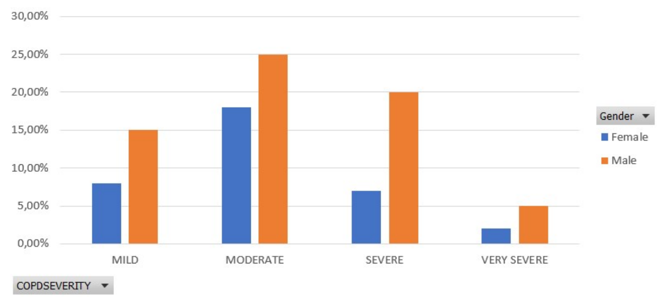

Figure 2: Participants distribution per COPD severity level and gender

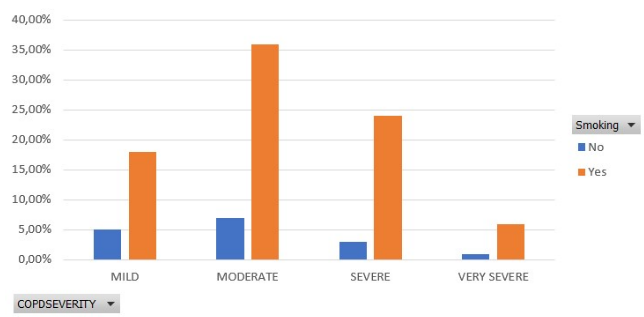

Figure 3: Participants distribution per COPD severity level and smoking status

four levels: mild, moderate, severe and very severe. One-against-one splits the multi-class classification problem into $K = 4$ binary classification sub-problems. Hence, the original dataset is divided into one dataset for each of the $K = 4$ classes versus every other individual class. Unlike one-against-one (OAO), the one-against-all (OAA) constructs $K = 4$ models where the n-th model of these is trained with all of the subjects in the n-th class with positive labels and the rest subjects with negative labels. These schemes are captured in Table 2. To solve the binary classification problems, Logistic Regression and Support Vector Machine models will be exploited and evaluated in the following sections.

OAA			OAO				
mild	moderate severe very severe	*1 binary classifier*	**mild**	moderate	severe	very severe	*K-1 binary classifiers*
moderate	mild severe very severe	*1 binary classifier*	**moderate**	-	very severe	moderate	*K-2 binary classifiers*
severe	mild moderate very severe	*1 binary classifier*	**very severe**	-	-	moderate	*K-3 binary classifiers*
very severe	mild moderate severe	*1 binary classifier*	-	-	-	-	-
Total		*K binary problems*	**Total**	-	-	-	*K(K-1)/2 binary problems*

Table 2: Multiclass Classification schemes OAO vs OAA in the COPD with $K = 4$

5 DATA PREPROCESSING AND FEATURE RANKING

In the context of this study, we applied random oversampling to generate a balanced dataset in terms of the target classes. This step aims to design a fair AI system whose discrimination ability will be similar to all patients irrespective of the COPD disease level.

The ML-based models were trained and tested considering all features described in Section 3 and the most important for COPD, according to the relevant literature. Also, we engineered two more features, the Age group (as shown in Figure 1) and CAT level (low, medium, high, very high) [15]. To evaluate the order of importance of an attribute f, the Gain Ratio (GR) method was employed. The Gain Ratio of an attribute f is estimated as $GR(f) = \frac{IG(f)}{H(f)}$, where $IG(f)$, $H(f)$ represent the Information Gain and entropy of f, respectively [23]. In the balanced dataset, the ranking of the features and the related values are as follows: FEV1PRED 0.6918, FEV 0.6504, MWT2 0.6198, MWT1Best 0.534, FVCPRED 0.5104, MWT1 0.4944, FVC 0.4825, CAT 0.3098, CAT LEVEL 0.267, SGRQ 0.2443, Muscular 0.2113, IHD 0.1445, Age group 0.1409, AtrialFib 0.0963, Hypertension 0.0657, Diabetes 0.0571, Gender 0.0158, Smoking 0.0105, Age 0 and PackHistory 0. The ranker results show that the features that concern the age and packet history don't contribute to the current data models. However, we maintained them since current models will be retrained with new data, and their importance will be re-investigated.

6 EVALUATION AND RESULTS

This section describes the performance of two ML models using 10-cross validation on the balanced dataset. The experiments were executed in the WEKA environment. To evaluate the validity of the data set used to the early predict the severity level of COPD in new patients, we examined the OAO and OAA methods using a multinomial LR model with a ridge estimator and an SVM. Concerning SVM, we exploited LibSVM [5] setting $\gamma = 0.01$ and using i) linear kernel (where $C = 0$, $d = 1$), ii) polynomial kernel (where $c = 2$, $d = 2$) and iii) RBF kernel with $C = 2$.

	OAA	OAO
linear	68%	83%
polynomial	81%	81%
rbf	73%	81%

Table 3: SVM models accuracy

To evaluate the models' classification performance, we based on the confusion matrix and, specifically, on precision, recall and accuracy, which are defined as:

$$Precision = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k}, Recall = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FN_k},$$
$$Accuracy = \frac{1}{K}\sum_{k=1}^{K}\frac{TN_k + TP_k}{TN_k + TP_k + FN_k + FP_k} \quad (1)$$

In Table 3, it is recorded the accuracy of the SVM models. As for the LR model, it achieved an accuracy of 77% and 79% for the OAO and OAA schemes, respectively. The accuracy of SVM for polynomial and rbf (radial basis function) kernels is 4% higher, while when using linear kernel it is 6% higher than the one of LR, in the OAO scheme. In the case of the OAA scheme, the accuracy of LR is 11% and 6% higher than that of SVM with linear and rbf kernel, correspondingly. However, its accuracy is 2% lower than SVM with a second-degree polynomial kernel.

Moreover, in Tables 4 and 5, we present the models' performance for each multi-class classification scheme, focusing on Precision, Recall and Area Under Curve (AUC). The higher the AUC, the more effective the model will be in distinguishing between the four COPD severity categories. Specifically, the AUC reveals that there is a high probability that the models will be able to distinguish the four categories. From the detailed comparison in terms of the SVM kernel type and the selected multi-class classification scheme, we see that OAO is a competitive approach. Similar conclusions are derived for the LR model. Considering all performance metrics outcomes, we conclude that SVM is more flexible, powerful and

Precision						Recall						AUC						Class
linear		polynomial		rbf		linear		polynomial		rbf		linear		polynomial		rbf		
OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	OAA	OAO	
0,727	0,778	0,840	0,778	1,000	1,000	0,640	0,840	0,840	0,840	0,720	0,720	0,884	0,920	0,787	0,923	0,871	0,914	**Mild**
0,471	0,826	0,727	0,750	1,000	0,568	0,320	0,760	0,640	0,720	0,440	1,000	0,769	0,899	0,809	0,886	0,767	0,864	**Moderate**
0,704	0,900	0,864	0,944	0,731	1,000	0,760	0,720	0,680	0,760	0,760	0,520	0,866	0,893	0,854	0,885	0,826	0,888	**Severe**
0,735	0,833	0.806	0,806	0,556	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0.,987	0,979	0,861	0,979	1,000	1,000	**Very Severe**

Table 4: Performance of SVM model with the OAA and OAO methods

Precision		Recall		AUC		Class
OAA	OAO	OAA	OAO	OAA	OAO	
0,826	0,741	0,760	0,800	0,849	0,921	**Mild**
0,692	0,609	0,720	0,640	0,819	0,883	**Moderate**
0,850	0,818	0,680	0,760	0,886	0,972	**Severe**
0,806	0,893	1,000	1,000	0,955	1,000	**Very Severe**

Table 5: Performance of LR model with the OAA and OAO methods

stable and thus a candidate model to be incorporated into the AI framework of the GATEKEEPER system.

7 CONCLUSIONS

The science of informatics and especially the fields of AI and ML play a key role in the early prediction of the COPD disease level besides various other chronic conditions that the elderly may suffer, such as high cholesterol, hypertension, diabetes, etc. This paper introduces a multi-class framework investigating the performance of two ML methods, Logistic Regression and Support Vector Machine. Both models presented high AUCs, a fact that seems promising for the discrimination ability of the models concerning the four severity stages of COPD. However, the SVM classifier achieved more consistent outcomes and generally superior performance.

In future work, we aim to investigate the relationship of attributes to each class by studying various attribute ranking and selection techniques to improve the models' efficiency. Finally, we will explore the capabilities provided by other models from Machine and Deep Learning.

ACKNOWLEDGMENTS

This work has been supported by the European Union's H2020 research and innovation programme GATEKEEPER under grant agreement No 857223, SC1-FA-DTS-2018-2020 Smart living homes-whole interventions demonstrator for people at health and social risks.

REFERENCES

[1] Davies Adeloye, Dhiraj Agarwal, Peter J Barnes, Marcel Bonay, Job F van Boven, Jamie Bryant, Gaetano Caramori, David Dockrell, Anthony D'Urzo, Magnus Ekström, et al. 2021. Research priorities to address the global burden of chronic obstructive pulmonary disease (COPD) in the next decade. *Journal of global health* 11 (2021).

[2] Sotiris Alexiou, Elias Dritsas, Otilia Kocsis, Konstantinos Moustakas, and Nikos Fakotakis. 2021. An approach for Personalized Continuous Glucose Prediction with Regression Trees. In *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM).* IEEE, 1–6.

[3] Åsa Athlin, Maaike Giezeman, Mikael Hasselgren, Scott Montgomery, Karin Lisspers, Björn Ställberg, Christer Janson, and Josefin Sundh. 2021. Prediction of Mortality Using Different COPD Risk Assessments–A 12-Year Follow-Up. *International Journal of Chronic Obstructive Pulmonary Disease* 16 (2021), 665.

[4] Kim Beernaert, Joachim Cohen, Luc Deliens, Dirk Devroey, Katrien Vanthomme, Koen Pardon, and Lieve Van den Block. 2013. Referral to palliative care in COPD and other chronic diseases: a population-based study. *Respiratory medicine* 107, 11 (2013), 1731–1739.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.

[6] Elias Dritsas, Sotiris Alexiou, Ioannis Konstantoulas, and Konstantinos Moustakas. 2022. Short-term Glucose Prediction based on Oral Glucose Tolerance Test Values. In *International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF*, Vol. 5. 249–255.

[7] Elias Dritsas., Sotiris Alexiou., and Konstantinos Moustakas. 2022. Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. In *Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE,*. INSTICC, SciTePress, 315–321.

[8] Elias Dritsas, Nikos Fazakis, Otilia Kocsis, Nikos Fakotakis, and Konstantinos Moustakas. 2021. Long-Term Hypertension Risk Prediction with ML Techniques in ELSA Database. In *International Conference on Learning and Intelligent Optimization.* Springer, 113–120.

[9] Nikos Fazakis, Elias Dritsas, Otilia Kocsis, Nikos Fakotakis, and Konstantinos Moustakas. 2021. Long-Term Cholesterol Risk Prediction with Machine Learning Techniques in ELSA Database. In *13th International Joint Conference on Computational Intelligence (IJCCI).* SCIPTRESS, 445–450.

[10] Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, and Konstantinos Moustakas. 2021. Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access* 9 (2021), 103737–103757.

[11] Tadahiro Goto, Carlos A Camargo Jr, Mohammad Kamal Faridi, Brian J Yun, and Kohei Hasegawa. 2018. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *The American journal of emergency medicine* 36, 9 (2018), 1650–1654.

[12] JoonNyung Heo, Jihoon G Yoon, Hyungjong Park, Young Dae Kim, Hyo Suk Nam, and Ji Hoe Heo. 2019. Machine learning–based model for prediction of outcomes in acute stroke. *Stroke* 50, 5 (2019), 1263–1265.

[13] Ali Hussain, Ikromjanov Kobiljon Komil Ugli, Beom Su Kim, Minji Kim, Harin Ryu, Satyabrata Aich, and Hee-Cheol Kim. 2021. Detection of Different stages of COPD Patients Using Machine Learning Techniques. In *2021 23rd International Conference on Advanced Communication Technology (ICACT).* IEEE, 368–372.

[14] Ioannis Konstantoulas, Otilia Kocsis, Elias Dritsas, Nikos Fakotakis, and Konstantinos Moustakas. 2021. Sleep Quality Monitoring with Human Assisted Corrections. In *International Joint Conference on Computational Intelligence (IJCCI).* SCIPTRESS, 435–444.

[15] Fernanda C Lanza, Rejane AS Castro, Anderson A de Camargo, Drielly JM Zanatta, Samia Rached, Rodrigo Athanazio, Alberto Cukier, Rafael Stelmach, and Simone Dal Corso. 2018. COPD Assessment Test (CAT) is a valid and simple tool to measure the impact of bronchiectasis on affected patients. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 15, 5 (2018), 512–519.

[16] Nancy K Leidy, Karen G Malley, Anna W Steenrod, David M Mannino, Barry J Make, Russ P Bowler, Byron M Thomashow, RG Barr, Stephen I Rennard, Julia F Houfek, et al. 2016. Insight into best variables for COPD case identification: a random forests analysis. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation* 3, 1 (2016), 406.

[17] Lixiang Li, Zihang Yang, Zhongkai Dang, Cui Meng, Jingze Huang, Haotian Meng, Deyu Wang, Guanhua Chen, Jiaxuan Zhang, Haipeng Peng, et al. 2020. Propagation analysis and prediction of the COVID-19. *Infectious Disease Modelling* 5 (2020), 282–292.

[18] Xu Min, Bin Yu, and Fei Wang. 2019. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Scientific reports* 9, 1 (2019), 1–10.

[19] Marc Miravitlles and Anna Ribera. 2017. Understanding the impact of symptoms on the burden of COPD. *Respiratory research* 18, 1 (2017), 1–11.
[20] Matthew Moll, Dandi Qiao, Elizabeth A Regan, Gary M Hunninghake, Barry J Make, Ruth Tal-Singer, Michael J McGeachie, Peter J Castaldi, Raul San Jose Estepar, George R Washko, et al. 2020. Machine learning and prediction of all-cause mortality in COPD. *Chest* 158, 3 (2020), 952–964.
[21] Anam Mustaqeem, Syed Muhammad Anwar, and Muahammad Majid. 2018. Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Computational and mathematical methods in medicine* 2018 (2018).
[22] Lokman Saleh, Hamid Mcheick, Hicham Ajami, Hafedh Mili, and Joumana Dargham. 2017. Comparison of machine learning algorithms to increase prediction accuracy of COPD domain. In *International Conference on Smart Homes and Health Telematics*. Springer, 247–254.
[23] Marwa Trabelsi, Nida Meddouri, and Mondher Maddouri. 2017. A new feature selection method for nominal classifier based on formal concept analysis. *Procedia Computer Science* 112 (2017), 186–194.