Discrete separation of patients' profiles for chronical obstructive pulmonary disease context-aware healthcare efficient systems

Hamid Mcheick[1], Farah Diab[2]
[1]Department of Computer Science and Mathematics, University of Quebec at Chicoutimi, Chicoutimi (QC), Canada
[2]Department of Computer Science-I, Ecole doctorale, Beirut, Lebanon

Article Info

ABSTRACT

According to the Public Health Agency of Canada (PHAC), the symptoms of chronic obstructive pulmonary disease (COPD) are shortness of breath, coughing, and sputum production. Many studies estimate that COPD will become the third-leading cause of death worldwide by 2030 (WHO, 2008). Pervasive healthcare systems cover healthcare issues, including chronic diseases; they help patients to manage their own health information and healthcare services at any time and in any place. We developed a COPD healthcare system based on a combination of the parameters of patients. The main goal is to avoid the severe phases of the disease by monitoring them. This combination of risk factors provides in total 600 profiles from data, with 88.5% accuracy. However, many studies have focused on and shown the issues of the effectiveness and accuracy of these systems. The problem is to apply a new classification model to detect the severe phases of the disease early. Therefore, instead of working on COPD parameters, we design and validate a profile-based classification model of patients. This model will facilitate the building of a rule-based framework. In addition, the accuracy of our extended COPD system is improved using the classification and separation of patients' profiles.

Corresponding Author:

Hamid Mcheick
Department of Compter Science and Mathematics, University of Quebec at Chicoutimi
555 Boul De l'Universite, G7H-2B1, Chicoutimi (Quebec), Canada
Email: mcheick_mcheick@uqac.ca

1. INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is one of the leading causes of death in the United States, affecting 16 million Americans (National Heart, Lung, and Blood Institute) [1] and 1.7 million Canadians. In terms of its death rate, one Canadian dies from COPD every hour, translating to 24 people per day and one-third of those who die from lung disease in this country (BREATHE, the lung association) [2]. The World Health Organization (WHO) announced that worldwide, COPD affected nearly 200 million people in 2016 and caused the death of 3.17 million people in 2015 [3]. In addition, the disease has high financial costs: The direct and indirect economic burden of the disease in Canada, for example, exceeds three billion dollars (PHAC). Telemedicine is evolving to protect patients and prevent chronic disease exacerbations by allowing the interaction of medical staff and patients without travelling by managing the disease status using telecommunication frameworks. In 2019, Ajami *et al.* [4] argued, "Recent years have witnessed a widespread increase in the number of telemedicine projects. This kind of intervention can open a window into the COPD patient's life to assist with self-management and prevent declines. Telehealth refers

to the remote monitoring and care of patients outside of the hospital setting. Typically, these systems are used for certain chronic diseases that are associated with frequent relapses. The role of telemedicine in COPD is still being discussed". Another study in 2019 that Gisler [5] conducted described the role of telehealth in lowering the number of hospital readmissions by providing services aimed at improving patients' quality of life: "However, research studies have shown that pulmonary rehabilitation (multidisciplinary services aimed at improving the quality of life in patients) managed to reduce readmissions by 56%".The telehealth domain is related to pervasive computing, self-adaptation, and contextual awareness for patients. In 2020, Mcheick *et al.* [6] presented architecture for a context-aware self-adaptive system that can be used to develop a COPD healthcare telemonitoring system. The system is backed out by a medical rules engine in the COPD domain that is used as the knowledge base to determine safe ranges for patient's biomarkers and external factors, then to detect the precluding actions needed to be taken to prevent severe exacerbations in the patient's health state.

Context-aware systems are an important topic in telehealth to reduce the risk of factors (context) of patients. Therefore, many studies have discussed the importance of these models. One of these studies (Kang *et al.* [7]) developed a context-aware framework that considered the ability of wearable sensors and middleware he althcare services to exchange information, while Oliveira *et al.* [8] proposed a decision-making framework for public healthcare systems: It is a context-aware framework called "LARISSA" that was proposed for telemonitoring inside the family's home. Kim *et al.* [9] proposed a pervasive ontology environment model that allows for extracting and classifying contextual information to implement healthcare services by considering medical references. Lo *et al.* [10] proposed a decision support system: The Ubiquitous Context-Aware Healthcare Service System (UCHS), which uses microsensors and integrates radio frequency identifier (RFID) to sense the user's vital signs, such as heart rate, respiratory rate, blood pressure, blood sugar, temperature, and includes an electrocardiogram. In addition, Mcheick *et al.* [11] built an ontology healthcare model based on the current context of patients, which made monitoring processes more accurate and proved the importance of user activity to define the context of medical application. Ajami *et al.* [4] proposed an ontology-based model to support ubiquitous healthcare systems for COPD patients by executing a sequential modular approach consisting of patients, disease, location, devices, activities, environment, and services to deliver personalized, real-time medical care for COPD patients. This project aims to set safe and dynamic boundaries for vital signs and assess environmental risk factors. This solution implements an interrelated set of ontologies with a logical base of semantic web rule language (SWRL). The rules are derived from the medical guidelines and expert pneumologists' definitions to handle all contextual situations. In 2019, Ajami *et al.* [4] presented validation for this proposition, where they explained the methods for extracting the medical rules of different contextual events. The research of Ajami *et al.* [4] examined the normal ranges of vital parameters during different activities of daily life and set a threshold for environmental conditions, whether indoors or outdoors, which was adapted to suit each patient's medical profile: The accuracy for vital signs was 89%. Moreover, only 600 profiles are extracted using rule-based classification and ontology. This model of Ajami *et al.* [4] had high accuracy, but it can be improved by extracting and separating large number of profiles. Since the last decade, machine learning has been applied in the medical domain for diagnostic, context-aware healthcare systems, self-management, and treatment and to improve the management of big and complex data to make predictions and prevent risks. Many researchers have studied the capacity of machine learning in the medical domain to predict diseases, define risk thresholds, and develop interactive healthcare systems [12]–[14]. Several context-aware systems focus on the relevant attributes of COPD exacerbations due to the large number of attributes that affect COPD risk factors. Furthermore, Himes *et al.* [15] identified clinical factors that modulate the risk of progression to COPD among asthma patients. As a result of this study, a model composed of age, sex, race, smoking history, and eight comorbidity variables can predict COPD in an independent set of patients with an accuracy of 83.3% using the Bayesian network. Moreover, Amalakuhan *et al.* [16] used the Random Forest model, the accuracy was 75% when dealing with attributes and the highest correlation, with a focus on hospital readmission. In 2012, Raghavan *et al.* [17] presented a model with an accuracy of 77% to identify patients at risk for COPD by combining eight components of the COPD assessment test (CAT) with smoking history and post-bronchodilator spirometry. Stepwise logistic regression analysis was applied to define the variables related to the presence of airway obstruction. In another study, Mcheick *et al.* [18] suggested a system called the helper context-aware engine system (HCES), which aims to help medical staff and patients by making correct decisions through selecting the most relevant contextual attributes and predicting exacerbations for patients using Naïve Bayes. It had high accuracy (80%) when selecting attributes. Moreover, a study by Mcheick *et al.* in 2017 [19] extended the existing HCES [19] using the Bayesian network for prediction, and accuracy was improved to 81.5%. Similarly, the models presented in these references [15]–[19] have moderate accuracies and, therefore, were limited in this performance, mainly because in the medical domain, it is important to use many classification techniques to improve these models. In 2019, Cavailles *et al.* [20] suggested an machine learning model for the identification of patients' profiles with a high risk of hospital

readmission for acute COPD exacerbations (AECOPD) in France to estimate the cost of re-hospitalization. However, this model did not take into account all parameters. It neglected the gold stage, smoking status, medication, and body mass index (BMI), but age, gender, and comorbidities were taken into consideration when classifying and applying the Decision Tree algorithm. In 2020, Vora *et al*. [21] also discussed COPD classification to predict COPD gold stages for patients using machine learning algorithms such as the support vector machine (SVM) and k-nearest neighbor (KNN). This model helps medical staff to predict the severity of patients' COPD but does not address risk factors by defining thresholds in relation to patients' profiles.

In this paper, we focus on the context-aware, rule-based system using our study of Ajami *et al*. [4] by introducing supervised machine learning classification and creating patient groups, called profiles. This separation of profiles improves classification accuracy and simplifies the building of a rule-based, context-aware system that combine multiple COPD parameters. Based on this analysis, this paper will answer the following questions:

−   How can we protect patients against risk factors?
−   How can we reduce the transition to a severe phase and disability using the context-aware, rule-based system that our research team developed [4]?

In particular, we focus on how we can improve classification accuracy to help patients and physicians by defining vital sign thresholds for each profile. Our goal is to monitor the progress of COPD to improve treatment effectiveness and reduce risk factors; specifically, this paper aims to apply machine learning classification algorithms to improve the accuracy of rule-based healthcare systems for COPD patients' statuses. Additionally, we design a classification model for COPD patients using profiles based on a combination of parameters. This model extracts the possible profiles using the separation of concerns technique.

In this research, two types of classifications are performed using Naïve Bayes and decision tree algorithms on large medical datasets after performing data preprocessing and data combination for parameters that experts have defined [4]. The profile classification aims to define rules after the prediction of the patient profile, knowing that these rules are defined based only on vital signs.

The contribution can be summarized in three points:

−   Apply machine learning algorithms to COPD using a large number of rules and data.
−   Combine COPD parameters into a large number of profiles based on what expert pneumologists [4] have defined, as well as on the separation of concerns.
−   Simplify the building of a rule-based framework using the discrete separation of concerns by classifying risk factors (parameters) in the profiles.

This paper is organized: section 2 describes the proposed method. section 3 describes the proposed classification model, while section 4 presents the results of the experiments and a comparison with our previous research study of Ajami *et al*. [4]. Finally, section 5 concludes this research and proposes future work.

2.   THE PROPOSED METHOD

In this research, we used the system as shown in Figure 1 composed of four layers proposed in [4], but with some modifications and extensions in the processing layer. We changed the ontology reasoning engine to the classification engine to separate patients according to their profiles to simplify the building of rules, decrease risk factors, and improve system accuracy.

Our proposed architecture consists of four layers. The first is the acquisition layer, which allows for collecting patients' data from different sources. This first layer can collect virtual and concrete data using existing databases and the internet of things (IoT) devices. Semantic formalization is often used to interpret complex information, which would make information meaningful and accessible to machines (second layer). This second layer uses the ontology as knowledge representation model in the most of the cases worldwide. The third layer is the principal layer for completing our proposition, and it contains the processing engine for data preprocessing combination and the classification engine that allows for the generation of rules according to profiles. This processing layer applies the algorithms of artificial intelligence to realize many tasks such as extract profiles and monitor patients (see details in section 3). The application layer is used for telemonitoring and risk assessment using the interfaces designed for physicians and users.

As mentioned before, this paper aims to enhance the accuracy of a rule-based, context-aware healthcare system, and we use the data used Ajami *et al*. [4]. This medical dataset contains 339 929 records with a high number of dimensions (58). We describe the medical dataset for COPD patients used in this research to apply our model. Therefore, this description can help us to understand the targeted data of the research. The medical dataset contains parameters of COPD when defining profiles, as experts have defined

[4], such as age, gender, smoking status, BMI, comorbidities (1, 2, 3), COPD gold stage, and medication. Other parameters include the COPD patient's historical vital signs during the levels of activity.
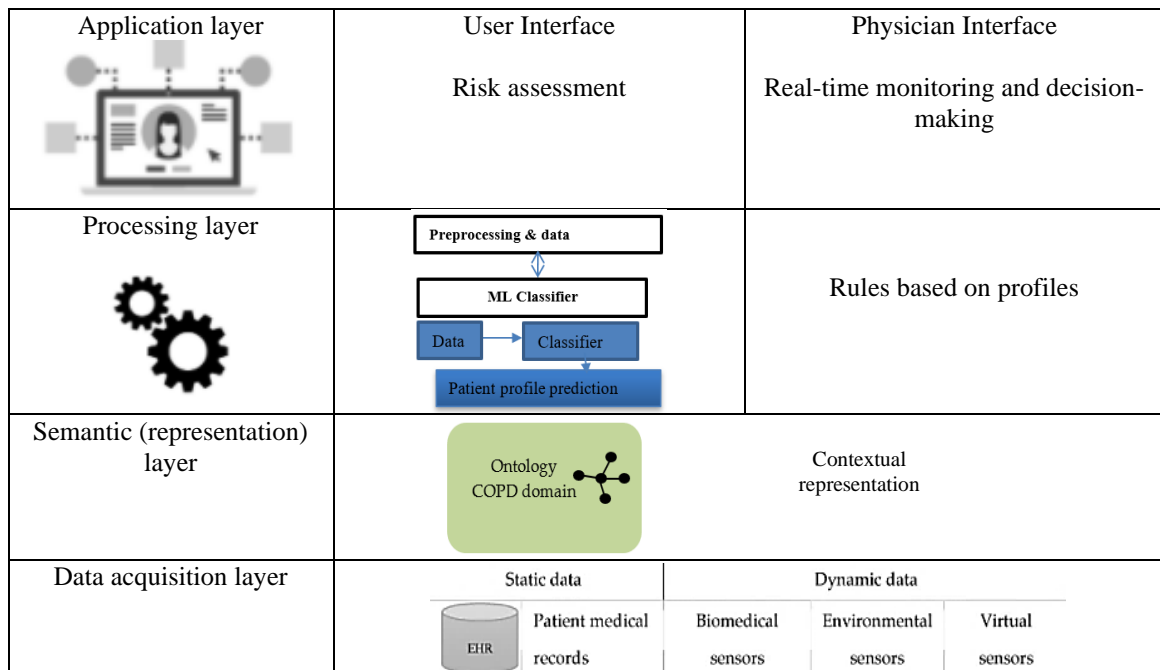
Application layer	User Interface Risk assessment	Physician Interface Real-time monitoring and decision-making
Processing layer	Preprocessing & data ML Classifier Data → Classifier Patient profile prediction	Rules based on profiles
Semantic (representation) layer	Ontology COPD domain	Contextual representation
Data acquisition layer	Static data EHR — Patient medical records	Dynamic data Biomedical sensors — Environmental sensors — Virtual sensors

Figure 1. Healthcare system architecture

3. METHOD: CONTEXT-AWARE MACHINE LEARNING CLASSIFICATION MODEL FOR COPD

In this section, we present the context-aware classification model to apply "classification" to the medical dataset. Our model consists of five phases: First, we start with data preparation in the "data preprocessing phase" to get a clean and correct form of the data for our medical dataset. The second phase is the "data combination phase" using the final form of the records, which will allow us to extract profiles in the "profiles extraction phase". We finish with the "classification phase", followed by the "definition of monitoring rules based on the profiles of the patients phase" as shown in Figure 2.

This model uses existing phases, such as data preprocessing in data science, classification application in machine learning, and rule-based algorithms in artificial intelligence, which were used when defining monitoring rules based on the profiles of patients. Our contribution is the data combination and profile extraction phases. Our model can be used for other healthcare systems, not only for COPD management.

Furthermore, our model consists of a workflow, which was composed:
− Data preprocessing
− Data combination
− Profile extraction
− The application of machine learning classification
− The definition of monitoring rules based on the profiles of patients

3.1. Data preprocessing

Geeksforgeeks [22] defined data preprocessing as "a data mining technique, which is used to transform the raw data in a useful and efficient format". During our data preparation or preprocessing, our goal was to have clean and efficient data while taking into consideration the sensitivity of patients' medical records. Any record with a lack of features was ignored, and we performed two steps: a) Data cleansing and b) Data transmission.
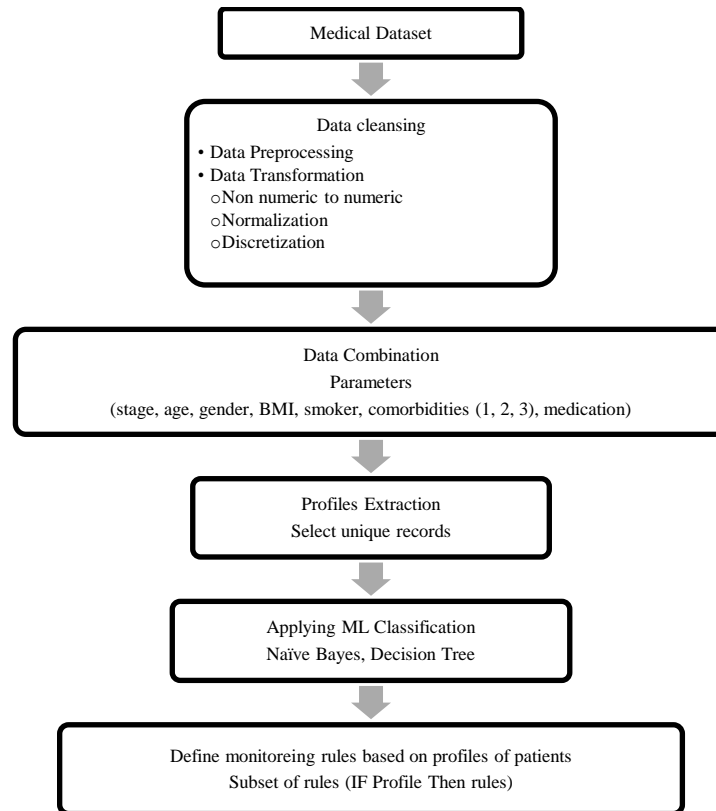
```
                    ┌─────────────────────────────────┐
                    │          Medical Dataset         │
                    └─────────────────────────────────┘
                                    │
                                    ▼
          ┌───────────────────────────────────────────────────┐
          │  Data cleansing                                    │
          │  • Data Preprocessing                              │
          │  • Data Transformation                             │
          │      ○Non numeric to numeric                       │
          │      ○Normalization                                │
          │      ○Discretization                               │
          └───────────────────────────────────────────────────┘
                                    │
                                    ▼
    ┌──────────────────────────────────────────────────────────────┐
    │                       Data Combination                        │
    │                          Parameters                           │
    │   (stage, age, gender, BMI, smoker, comorbidities (1, 2, 3),  │
    │                          medication)                          │
    └──────────────────────────────────────────────────────────────┘
                                    │
                                    ▼
          ┌───────────────────────────────────────────────────┐
          │               Profiles Extraction                  │
          │               Select unique records               │
          └───────────────────────────────────────────────────┘
                                    │
                                    ▼
          ┌───────────────────────────────────────────────────┐
          │            Applying ML Classification              │
          │            Naïve Bayes, Decision Tree              │
          └───────────────────────────────────────────────────┘
                                    │
                                    ▼
      ┌───────────────────────────────────────────────────────────┐
      │   Define monitoreing rules based on profiles of patients   │
      │         Subset of rules (IF Profile Then rules)            │
      └───────────────────────────────────────────────────────────┘
```

Figure 2. Machine learning classifier model

3.1.1. Data cleansing

In our study, data cleansing was performed to handle missing data, so the first procedure was to remove records with attributes that had null or undefined values as shown in Figure 3; for example, we did not fill in missing values based on the average values because the data was highly sensitive and related to the patient's life. As a result of the data cleansing, we got 331 462 records out of the 339 889 records (8427 records were omitted). Before starting this step, we removed three features: height (cm), height (in), and weight (kg) because they were defined by the BMI feature (BMI = weight (kg) / [height (m)]^2). In addition, we deleted the baseline temperature in Fahrenheit because it was a duplication of the data—we used only the Celsius metric.

Current FEV1	Baseline VO2	VO2 Reserve	VO2 max	VO2 During light exercis	VO2 During moderate exercise	VO2 During vigorous exercise
0,345 L	1.529	1.782903359	3.311903359	1,897 ml/kg/min	2,281 ml/kg/min	2,984 ml/kg/min
1,878 L	2.751	18.21451189	20.96551189	7,302 ml/kg/min	10,806 ml/kg/min	16,838 ml/kg/min
1,845 L	1.891	11.30434351	13.19534351	4,419 ml/kg/min	8,07 ml/kg/min	10,543 ml/kg/min
0,733 L	2.526	4.44706286	6.97306286	3,507 ml/kg/min	4,919 ml/kg/min	5,495 ml/kg/min
2,96 L	2.801	14.33384605	17.13484605	5,212 ml/kg/min	9,25 ml/kg/min	11,951 ml/kg/min
2,162 L	3.057	11.43977173	14.49677173	6,284 ml/kg/min	8,884 ml/kg/min	12,137 ml/kg/min
1,564 L	2.537	13.67784704	16.21484704	4,778 ml/kg/min	9,056 ml/kg/min	11,509 ml/kg/min
1,034 L	1.735	5.619206618	7.354206618	3,394 ml/kg/min	4,921 ml/kg/min	6,24 ml/kg/min
0,563 L	2.573	5.7924801	8.3654801	3,778 ml/kg/min	5,569 ml/kg/min	6,382 ml/kg/min
1,808 L	2.903	13.67674557	16.57974557	7,002 ml/kg/min	9,443 ml/kg/min	12,564 ml/kg/min
1,861 L	1.708	20.0792658	21.7872658	5,513 ml/kg/min	11,984 ml/kg/min	17,602 ml/kg/min
1,108 L	1.809	18.53930817	20.34830817	4,211 ml/kg/min	9,335 ml/kg/min	14,218 ml/kg/min
1,147 L	1.911	9.099547194	11.01054719	3,302 ml/kg/min	6,765 ml/kg/min	9,543 ml/kg/min
0,647 L	2.28	0.074442573	2.354442573	2,297 ml/kg/min	2,32 ml/kg/min	2,339 ml/kg/min
1,196 L	1.837	10.91955555	12.75655555	3,458 ml/kg/min	6,626 ml/kg/min	9,628 ml/kg/min
2,244 L	1.965	12.75666372	14.72166372	4,863 ml/kg/min	7,651 ml/kg/min	10,225 ml/kg/min
0,758 L	2.291	-0.46958617	1.821413834	#NUM!	#NUM!	#NUM!
1,059 L	2.694	4.836813294	7.530813294	3,602 ml/kg/min	5,359 ml/kg/min	6,566 ml/kg/min
1,102 L	1.302	4.848695341	6.150695341	1,919 ml/kg/min	3,522 ml/kg/min	4,925 ml/kg/min
0,77 L	2.006	5.950684443	7.956684443	3,475 ml/kg/min	4,618 ml/kg/min	5,699 ml/kg/min

Figure 3. Example of deletion

3.1.2. Data transformation

Data transformation was applied to transform the data into a meaningful and legible form to be used in classification. That is, it helped to make the data that the machine learning algorithms produced legible because they were math-based. Therefore, we replaced categorical values with numerical values. Table 1 uses the numberical values of the stages. Table 2 transforms the BMI values in a digital value. Table 3 converts medication correspondent values into a numerical value. Table 4 replaces comorbidities in numerical values as shown in Tables 1-4.

Table 1. COPD gold stage correspondent values (stages)

Stage	Value
Stage 1	1
Stage 2	2
Stage 3	3
Stage 4	4

Table 2. BMI correcpondent values (BMI)

BMI	Rang of BMI	Value
Underweight	<18.5	1
Healthy weight	18.5 to 24.9	2
Overweight	25 to 29.9	3
Obese	30 or higher	4

Table 3. Medication correspondent values (COPD Medication)

Formula	Value
LABA + SABA prn	1
LAAC + ICS + SABA prn	2
LAAC + ICS + SABA prn + OCS	3
LAAC + ICS + SABA prn + PDE4	4
LAAC + ICS + SABA prn + Theophylline	5
LAAC + ICS + SABA prn + Theophylline + PDE4	6
LAAC + ICS + SABA prn + Theophylline	7
LAAC + ICS + SABA prn + Theophylline + OCS	8
LAAC + ICS + SABA prn + Theophylline	9
LAAC + LABA + SABA prn	10
LAAC + LABA + SABA prn + PDE4	11
LAAC + LABA + SABA prn + Theophylline + OCS	12
LAAC + LABA + SABA prn + Theophylline + PDE4	13
LAAC + LABA + SABA prn +Theophylline	14
LAAC + LABA + SABA prn + Theophylline	15
LAAC + LABA + SABA prn + Theophylline	16
LAAC + SABA prn	17
LABA + Short-acting bronchodilator prn + Theophylline	18
SAAC + SABA	19
Short-acting bronchodilator prn	20

Table 4. Comorbidity correspondent values (Comorbidities)

Comorbidity	Value
Acid reflux	1
Anemia	2
Asthma	3
Chronic kidney	4
Congestive heart failure	5
Coronary artery	6
Diabetes	7
Dyspnea	8
High blood pressure	9
Pulmonary hypertension	10
None	0

The second step was data normalization, which involved applying 0 and 1 values for the gender and smoking status parameters:
− Male: 1; Female: 0
− Smoker: 1; non-smoker: 0

Finally, we applied discretization to the age feature in Table 5. This discretization is based on the methods used by the experts (doctors). The intervals are used to identify the profiles of patients to apply different rules of these profiles. This discretization is given in Table 5 given.

Table 5. Age correspondent values

Value	Age Range
1	40-50
2	50-60
3	60-70
4	70-80
5	Greater than 80

Table 6 contains the final form of the COPD parameters (features), such as gold stage, gender, age, BMI, smoking status, and comorbidity (1, 2 and 3). Experts (medicins) use these relevant parameters to handle the COPD disease. These parameters are relevant because they can reveal the exacerbations of patients.

Table 6. Final-form parameters

Gold Stage	Gender	Age	BMI	Smoking Status	Comorbidity 1	Comorbidity 2	Comorbidity 3	Medication
1	0	1	1	0	0	0	0	18
1	0	1	1	0	0	0	0	19
1	0	1	1	0	0	0	4	20
1	0	1	1	0	0	0	8	20
1	0	1	1	0	0	2	10	19
1	0	1	1	0	0	4	0	19

3.2. Data combination

Our main goal was to extract those profiles from the medical dataset that facilitated the patients' classification or grouping. Therefore, we combined the COPD parameters that the experts defined in [4] (stage, gender, smoking status, age, BMI, comorbidities (1, 2, 3), medication). Moreover, data combination consists of taking values to be used in the combination from the dataset. Furthermore, we could not apply the combination to all of the existing values of the parameters because when a combination had no records, it was automatically omitted from the study since we only worked with real or occurring values as shown in Table 7.

Table 7. Combination examples

Gold Stage	Gender	Smoker	Age	BMI	Comorbidity 1	Comorbidity 2	Comorbidity 3	Medication	CONCAT Combination
1	0	0	47	1	0	0	0	18	1–0–0–47–1–0–0–0–18
1	0	0	48	1	0	0	0	19	1–0–0–48–1–0–0–0–19
1	0	0	43	1	0	0	0	19	1–0–0–43–1–0–0–0–19
1	0	0	44	1	0	0	0	19	1–0–0–44–1–0–0–0–19
1	0	0	49	1	0	0	4	20	1–0–0–49–1–0–0–4–20

To understand the combination of profiles, consider a patient with the following values:
Stage: Stage 1 took 1
Gender: Female took 0
Age: 45 took 1 (between 40 and 50 when we use the experts' proposal in [4] or 45 when ignoring the experts' proposal)
BMI: 18 took 1
Smoker: No took 0
Comorbidities 1: None took 0
Comorbidities 2: Anemia took 2
Comorbidities 3: Pulmonary hypertension took 10
Medication: SAAC + SABA took 19
Therefore, the combination of values would be:

Stage 1–female–45–18–non-smoker–none–anemia–pulmonary hypertension–SAAC + SABA
After applying the parameters in the transformation phase, the values would be 1–0–1–1–0–0–2–10–19, which defines the combination of values to be used for profile extraction.
Each unique combination comprised a group of patients that had the same parameters, for example:
Patient X: combination 1–0–0–47–1–0–0–0–18 (with each number corresponding to a parameter value).
Patient Y: combination 1–0–0–47–1–0–0–0–18 (which is the same combination as Patient X).

3.3. Profile extraction process

We aimed to make the building of the rule-based system that Ajami *et al*. [4] proposed, which extracted 600 profiles, more efficient and simple. We designed a new model as shown in Figure 4 using the separation of profiles technique to extract and separate the most important profiles and maximize the number of profiles by applying data combination only for those parameters defined by expert in [4] (stage, gender, smoking status, age, BMI, comorbidities (1, 2, 3), medication). Then, we merged their values to extract unique combined values for defining a single profile. Finally, we selected all of the combinations with unique values, taking into consideration that many patients had the same profile because they had the same parameter combination. After combining the parameters, we extracted the patients' profiles by selecting unique combinations. Each unique combination was assigned a specific number, following which we set a profile number for each patient as shown in Table 8.
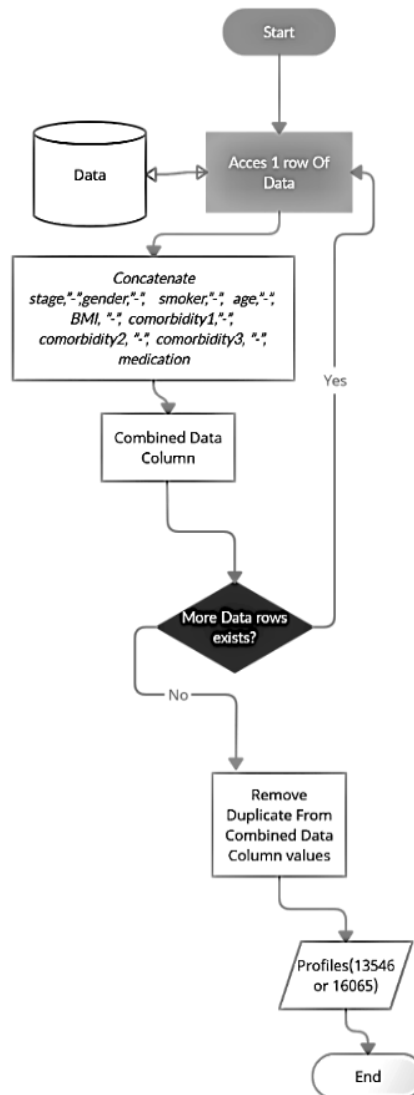
Figure 4. The algorithm used to identify the profiles of patients by combining the nine parameters
Table 8. Profile assignment

Profile	Gold Stage	Gender	Age	BMI	Smoking Status	Comorbidity 1	Comorbidity 2	Comorbidity 3	Medication
1	1	0	1	1	0	0	0	0	18
2	1	0	1	1	0	0	0	0	19
3	1	0	1	1	0	0	0	4	20
4	1	0	1	1	0	0	0	8	20
5	1	0	1	1	0	0	2	10	19
		0	1	1	0	0	4	0	19

The profiles as shown in Table 9 were the classes to be predicted, but there was a multiclass classification problem because it was more complicated than binary classification [23]. We identified 600 profiles as a simple model to extract more than twenty thousand rules [4]. Two classifications are proposed: a) Based on the expert's suggestion and b) Based on the combination of the all the factors.

Table 9. Extraction of profile

Expert's suggestion	Expert's suggestion neglected
13546 Profiles	16065 Profiles

3.4. Classification algorithms

Our proposal classified patients' profiles using a multiclass classification. Therefore, two algorithms were used to accomplish this step: Naïve Bayes and decision tree. These algorithms are applied and evaluated.

3.4.1. Naïve Bayes classifier

This task was accomplished using the Python language and multiple libraries (Pandas, Scikit-learn, Numpy) [24]. This was a multiclass classification problem because we had more than two classes; it was slightly challenging to deal with this type of problem. Here we had two proposals, one defined by the experts in [4], and the other similar to the experts' suggestion but without the discretization of the age feature. Naïve Bayes uses the Bayes theorem and is very reliable when using large data lengths. It is also simple and highly accurate [24].
Classification based on the experts' proposal for age categories:
− We had 13 546 classes (profile 1, profile 2 … profile 13 546).
− The training set was for 70% of the data as shown in Figure 5.
− The test set was for 30% of the data as shown in Figure 5.
− The Scikit-learn package was used to obtain the results (accuracy: metrics.accuracy_score).
− The performance of the model (accuracy) was 0.964 (96.4%).
Classification based on ignoring the experts' proposal for age categories:
− We had 16 056 classes (profile 1, profile 2 … profile 16 056).
− The training set was for 70% of the data as shown in Figure 5.
− The test set was for 30% of the data as shown in Figure 5.
− The Scikit-learn package was used to obtain the results (accuracy: metrics.accuracy_score).
− The performance of the model (accuracy) was 0.954 (95.4%).

3.4.2. The decision tree classifier

Decision Tree classification is a flowchart algorithm and is a simple algorithm with high performance that is easy to understand. It is composed of nodes, and each node represents the conditions of the data features. Leaf nodes represent the results or classes after going through the tree [25]. Like Naïve Bayes, decision tree is highly accurate, and it has a very high processing speed. Therefore, it was the second algorithm used to classify our data into patient profiles using the same parameters and proposal described in section 3.4.1.
Classification based on the experts' proposal for age categories:
− We had 13 546 classes (profile 1, profile 2 … profile 13 546).
− The training set was for 70% of the data as shown in Figure 5.
− The test set was for 30% of the data as shown in Figure5.
− The Scikit-learn package was used to obtain the results (accuracy: metrics.accuracy_score).
− The performance of the model (accuracy) was 0.963 (96.3%).

Classification based on ignoring the experts' proposal for age categories:
−   We had 16 056 classes (profile 1, profile 2 … profile 16 056).
−   The training set was for 70% of the data as shown in Figure 5.
−   The test set was for 30% of the data as shown in Figure 5.
−   The Scikit-learn package was used to obtain the results (accuracy: metrics.accuracy_score).
−   The performance of the model (accuracy) was 0.955 (95.5%).
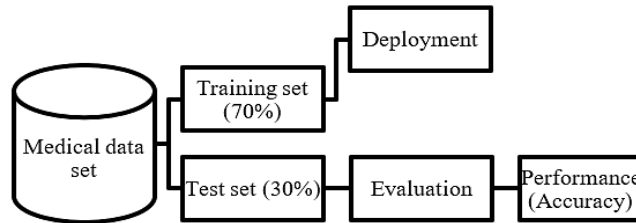
Figure 5. Classfication workflow

3.5. Defining the monitoring rules based on the profiles of patients

After the classification according to profiles, a subset of rules need to be defined for ensuring the accuracy of monitoring of patients' health by focusing on their vital signs, with or without activities that would trigger an alarm in case of risk (a severe phase of the disease). Based on the profiles, a rule-based system is a structure that allows for defining rules or conditions using profiles and used If-Then statements: If event A happens, then do an action. It was a set of predefined rules used to decide in case of the rule being satisfied. In our proposal, we highlighted some examples of predefined rules after defining the patients' profiles,

---

If Profile = 12
{HR during vigorous exercise}
then
if HR >144 then
{alarm should be triggered}
If Profile = 8
{Temperature (°C) as T during light exercise}
Then
if T > 36.7868
then
{alarm should be triggered}
If Profile = 270
{SpO$_2$ (Blood oxygen saturation) during vigorous exercise}
then

---

if SpO$_2$ > 92.25621053
then
{alarm should be triggered}
If Profile = 13 546
{RR (respiration rate) during light exercise}
then
if RR > 35.60676219
then
{alarm should be triggered}
If Profile = 3678
{PaO$_2$ (partial pressure of oxygen) during moderate exercise}
then
if PaO$_2$ > 89
then
{alarm should be triggered}

Therefore, the maximum number of profiles for the experiment shown in section 4 indicates that the accuracy increased and performance was enhanced. This is because the classification of patients was based on their profiles. When the medical record changed, the thresholds were predefined without any newly built rules because we took only the number of profiles (if profile, then threshold).

4.   RESULTS AND DISCUSSION

In our study, and as shown in Table 10, the accuracy improved by approximately 7% using Naïve Bayes and decision tree in comparison with the latest study of Ajami *et al*. [4], which proposed using a rule-based ontology framework for COPD patients. In their study [4], 600 profiles were extracted, and the accuracy was 88.5%. Therefore, the added value of this approach was the increase in accuracy level using machine learning classification and data combination. However, we think that 600 profiles are not sufficient to improve accuracy. Knowing this, when we used the "age" parameter without categorization (the age value was used, for example, 40, 41, 42…), the accuracy was less than when using "age" with discretization (when age was grouped into categories, for example, 40–50, 50–60, and so on, and we assigned a discrete value for each category of age).

As a result, it is important to use classification and data combination by extracting the most important profiles to improve rule-based context-aware systems for COPD patients. These rules protect their lives and get reliable healthcare systems that consider the sensitivity of medical data. The different algorithms of artificial intelligence show the accuracy obtained with different number of profiles. The algorithms will be evaluated with huge number of rules and can be improved by decomposing them into different categories [6]. In addition, In addition, this preliminary result needs to be evaluated with a huge number of data and rules, and by dividing the rules and services in different software units (modules) using our model [6].

Table 10. Accuracy comparison using the same data of patients

Algorithms	Number of profiles	Accuracy (%)
Naïve Bayes	13 546	96.4
Naïve Bayes	16 056	95.4
Decision Tree classification	13 546	96.3
Decision tree Classification	16 056	95.5
Rule-Based Classification [4]	600	89

5.   CONCLUSION

In this paper, the problem is to detect the severe phases of the COPD disease. Therefore, instead of working on COPD parameters, we design and validate a profile-based classification model of patients. We proposed an extension of context-aware healthcare system for COPD by using machine learning, supervised learning classification algorithms to identify and combine discrete patients' profiles. The results show that Naïve Bayes and decision tree are the most accurate for combining nine COPD parameters. This combination allows the extraction of huge number of profiles instead of only six hundred profiles. These combinations of profiles increase the accuracy of the prediction of exacerbations but they increase the number of rules, which need more execution time. The physiciens advice to use the profil of a patient with nine parameters decribed in this paper. When increasing the number of profiles by dividing them based on the advice of experts in the medical domain and using the machine learning algorithms of Naïve Bayes and decision tree, the accuracy is increased by seven per cent compared to the our accuracy of the rule-based classification. However, we should take into consideration the problem of imbalanced datasets. Therefore, resampling should be avoided due to the sensitivity of the data, and it is important to collect more data globally for experimenting machine learning and offer a reliable model to protect patients from risks. In future, we will work on making the study more applicable to real-life scenarios and collect more and global data to resolve the issues related to imbalanced datasets. The execution time will be evaluated with huge number of rules and can be improved by decomposing them into different categories and modules. In addition, we will apply deep learning algorithms to predictions, as well as big data technologies and real-time processing.

and F.D designed the model and analyzed the computational framework of the ontology. H.M. wrote the manuscript in discussion with F. D. Authors discussed the results and contributed to the final manuscript.

REFERENCES

[1]   L. National Heart and B. I. (NHLBI), "What is copd?," [Online]. Available: https://www.nhlbi.nih.gov/health-topics/copd. (Accessed: June 3, 2020).

[2]   l. a. L. F. Sheet, "The lung association, lung fact sheet," 2015, [Online]. Available: https://sk.lung.ca/aboutus/news-room/backgrounders-and-information-sheets/lung-fact-sheet. (Accessed: June 5, 2020).

[3]   G. A. Dumitrascu, "Chronic obstructive pulmonary disease (COPD)," *Decision Making in Anesthesiology: An Algorithmic Approach: Fourth Edition*, pp. 100–101, 2007, doi: 10.1016/B978-0-323-03938-3.50039-7.

[4]   H. Ajami, H. McHeick, and K. Mustapha, "A pervasive healthcare system for COPD patients," *Diagnostics*, vol. 9, no. 4, 2019, doi: 10.3390/diagnostics9040135.

[5]   S. Gisler, "Fewer copd patients readmitted after video rehabilitation, study says," 2019, [Online]. Available: https://copdnewstoday.com/2019/05/17/real-time-video-based-pulmonary-rehabilitation-lowershospital-readmission-rates-after-copd-exacerbation-study.

[6]   H. McHeick and J. Sayegh, "A self-adaptive and efficient context-aware healthcare model for copd diseases," *Informatics*, vol. 8, no. 3, 2021, doi: 10.3390/informatics8030041.

[7]   D. O. Kang, H. J. Lee, E. J. Ko, K. Kang, and J. Lee, "A wearable context aware system for ubiquitous healthcare," *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, pp. 5192–5195, 2006, doi: 10.1109/IEMBS.2006.259538.

[8]   M. Oliveira *et al.*, "A context-aware framework for health care governance decision-making systems: A model based on the Brazilian digital TV," *2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks", WoWMoM 2010 - Digital Proceedings*, 2010, doi: 10.1109/WOWMOM.2010.5534979.

[9]   J. Kim and K. Y. Chung, "Ontology-based healthcare context information model to implement ubiquitous environment," *Multimedia Tools and Applications*, vol. 71, no. 2, pp. 873–888, 2014, doi: 10.1007/s11042-011-0919-6.

[10]   C. C. Lo, C. H. Chen, D. Y. Cheng, and H. Y. Kung, "Ubiquitous healthcare service system with context-awareness capability: Design and implementation," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4416–4436, 2011, doi: 10.1016/j.eswa.2010.09.111.

[11]   H. McHeick, H. Ajami, and Z. Elkhaled, "Survey of health care context models and prototyping of healthcare context framework," *Simulation Series*, vol. 48, no. 9, pp. 422–429, 2016, doi: 10.22360/summersim.2016.scsc.072.

[12]   S. J. Hickey, "Naive Bayes classification of public health data with greedy feature slection," *Communications of the IIMA*, vol. 13, no. 2, pp. 87–98, 2013.

[13]   M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[14]   M. R. Ullah, M. A. R. Bhuiyan, and A. K. Das, "IHEMHA: Interactive healthcare system design with emotion computing and medical history analysis," *2017 6th International Conference on Informatics, Electronics and Vision and 2017 7th International Symposium in Computational Medical and Health Technology, ICIEV-ISCMHT 2017*, vol. 2018-Janua, pp. 1–8, 2018, doi: 10.1109/ICIEV.2017.8338606.

[15]   B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of chronic obstructive pulmonary disease (COPD) in Asthma patients using electronic medical records," *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 371–379, 2009, doi: 10.1197/jamia.M2846.

[16]   B. Amalakuhan, L. Kiljanek, A. Parvathaneni, M. Hester, P. Cheriyath, and D. Fischman, "A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem," *Journal of Community Hospital Internal Medicine Perspectives*, vol. 2, no. 1, p. 9915, 2012, doi: 10.3402/jchimp.v2i1.9915.

[17]   N. Raghavan *et al.*, "Components of the COPD assessment test (CAT) associated with a diagnosis of COPD in a random population sample," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 9, no. 2, pp. 175–183, 2012, doi: 10.3109/15412555.2011.650802.

[18]   H. Mcheick, L. Saleh, H. Mili, and H. Ajami, "HCES: Helper context engine system to predict relevant state of patients in COPD domain using Naïve Bayesian," *ACM International Conference Proceeding Series*, 2017, doi: 10.1145/3109761.3109792.

[19]   H. Mcheick, L. Saleh, H. Ajami, and H. Mili, "Context relevant prediction model for COPD domain using Bayesian belief network," *Sensors (Switzerland)*, vol. 17, no. 7, 2017, doi: 10.3390/s17071486.

[20]   A. Cavailles *et al.*, "Identification of patient profiles with high risk of hospital re-admissions for acute COPD exacerbations (AECOPD) in France using a machine learning model," *International Journal of COPD*, vol. 15, pp. 949–962, 2020, doi: 10.2147/COPD.S236787.

[21]   S. Vora and S. Chintan, "COPD classification using machine learning algorithms," *International Research Journal of Engineering and Technology*, vol. 6, pp. 608–611, 2008, [Online]. Available: www.irjet.net.

[22]   S. García, J. Luengo, and F. Herrera, "Data preprocessing in data mining," *Intelligent Systems Reference Library*, vol. 72, 2015, [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-in-datamining/: :text=Steps Involved in Data Preprocessing%3A 1 Data Cleaning%3A.

[23]   Y. Ahuja and S. Kumar Yadav, "Multiclass classification and support vector machine," *Global Journal of Computer Science and Technology Interdisciplinary*, vol. 12, no. 11, pp. 14–19, 2012, [Online]. Available: https://globaljournals.org/GJCST_Volume12/2-Multiclass-Classification-and.pdf.

[24]   A. Navlani, "Naïve bayes classifier tutorial: With python scikit-learn," 2018, [Online]. Available: https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn. (Accessed: July 18, 2020).

[25]   A. Navlani, "Python decision tree classification with scikit-learn decision tree classifier-data camp," 2018, [Online]. Available: https://www.datacamp.com/tutorial/decision-tree-classification-python%0Ahttps://www.datacamp.com/tutorial/decision-tree-classification-python%0Ahttps://www.datacamp.com/community/tutorials/decision-tree-classification-python. (Accessed: July 25, 2020).

BIOGRAPHIES OF AUTHORS

Professor Hamid Mcheick (ID) [g] [SC] [C] is a full professor in Computer Science department at the University of Quebec at Chicoutimi, Canada. He has more than 20 years of experience in both academic and industrial area. He has done his PhD in Software Engineering and Distributed System in the University of Montreal, Canada. He is working on design and adaptation of distributed and smart software applications; designing healthcare and IoT frameworks. He has supervised many post-doctorate, PhD, master and bachelor students. He has nine book chapters, more than 60 research papers in international journals and more than 140 research papers in international/national conference and workshop proceedings in his credit. Dr. Mcheick has given many keynote speeches and tutorials in his research area, particularly in healthcare systems, Pervasive and Ubiquitous computing, Distributed Middleware Architectures, Software Connectors, Internet of Things (IoT), Mobile Edge Computing, Fog Computing, and Cloud Computing. Dr. Mcheick has gotten many grants from governments, industrials and academics. He is a chief in editor, chair, co-chair, reviewer, member in many organizations (such as IEEE, ACM, Springer, Elsevier, Inderscience) around the world. He can be contacted at email: hamid_mcheick@uqac.ca.

Farah Diab (ID) [g] [SC] [C] received his Msc Master in computer science at Lebanese University, Lebanon. He has a long period of experience in software engineering area. Farah has a good experience in IT developer and monitoring and evaluation officer. He can be contacted at email: farah.a.diab@gmail.com.