Prediction of COVID-19 Diagnosis Based on OpenEHR Artefacts

Daniela Oliveira
University of Minho
Diana Ferreira
University of Minho
Nuno Abreu
Centro Hospitalar do Porto
Pedro Leuschner
Centro Hospitalar do Porto
António Abelha
University of Minho
José Machado (jmac@di.uminho.pt)
University of Minho

Research Article

Keywords: COVID-19, scalable data structure methodologies, SARS-CoV-2, openEHR architecture, algorithms

Posted Date: October 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-907764/v1

Prediction of COVID-19 Diagnosis based on OpenEHR Artefacts

Daniela Oliveira[1], Diana Ferreira[1], Nuno Abreu[2], Pedro Leuschner[2], António Abelha[1], and José Machado[1*]

[1]Algoritmi Research Center, University of Minho, Campus of Gualtar, Braga 4710, Portugal
[2]Centro Hospitalar Universitário do Porto, Porto 4099, Portugal
[*]Corresponding author: jmac@di.uminho.pt

ABSTRACT

The complexity and momentum of monitoring COVID-19 patients calls for the usage of agile and scalable data structure methodologies. A system for tracking symptoms and health conditions of suspected or confirmed SARS-CoV-2 infected patients was developed based on the openEHR architecture. All data on the evolutionary status of patients in home care as well as the results of their COVID-19 test were used to train different ML algorithms, with the aim of developing a predictive model capable of identifying COVID-19 infections according to the severity of symptoms identified by patients. The results obtained were promising, with the best model achieving an accuracy of 96.25%, a precision of 99.91%, a sensitivity of 92.58%, and a specificity of 99.92%, using the Decision Tree algorithm and the Split Validation method.

Introduction

Today's Health Information Systems (HIS) include numerous types of software, resulting in a wide range of versions and technologies employed, even within the same organization. Because of the lack of national and institutional guidelines, different parts of a given HIS represent the same information in different ways, providing a significant barrier to semantic interoperability. Information Models (IMs) capable of decreasing the barriers developed through time to achieve HIS interoperability are becoming more crucial. To accomplish this, existing global standards for the development of consistent and interoperable IMs are becoming increasingly important. Such standards can cover demographic, clinical, and administrative modules, as well as information access control.

All these issues compromise the potential of Information Technology (IT) to help daily clinical practice and knowledge production, and limits the ability to deploy reliable Clinical Decision Support Systems (CDSSs)[1,2]. Clinical data not only enables decisions for continuity of care, but also serves several secondary uses, such as medical and academic research, business intelligence indicators, and the discovery of new knowledge using Artificial Intelligence (AI) and ML (ML) algorithms. Although secondary, these usages are considered critical for continuous improvement in the delivery of quality healthcare. Hence, it is necessary to provide support for new data sources, including wearables, patient reported data, and external systems.

Therefore, it is of uttermost importance that distinct data representations are transformed and integrated according to a common data model[3–5].

COVID-19 is the name given by the World Health Organization (WHO) to the infectious disease caused by SARS-CoV2, the most recently discovered coronavirus, which is infecting large numbers of people worldwide and for which no country was prepared. As a result, during the COVID-19 pandemic, many countries' healthcare systems became overburdened. This was due to a variety of factors, including a lack of human and material capital while the demand for healthcare was increasing. Furthermore, the difficulties encountered in exchanging data in regular day-to-day work were exacerbated by the pandemic's pressure, putting even more strain on health professionals. Medical data is a valuable resource in these emergency situations, not only for clinical decisions but also for health political governance, since it provides up-to-date and real-time information on the pandemic's progression.

An ongoing pilot project focused on the OpenEHR specification was used by Centro Hospitalar Universitário do Porto (CHUP) institution to refine the COVID-19 patient's treatment workflow, which is presented on Figure 1. As a result, a hybrid solution that interacted with existing LSs was developed to ensure that users could communicate quickly and effectively with minimal effort.

A web application has been created to keep track of patients at home care. The data entry of this application was guaranteed through a form based on a template modeled in openEHR's clinical methodology. Each patient was free to report their symptoms and health status as many times as they wanted, either before or after learning the results of their SARS-CoV-2 test[6,7].

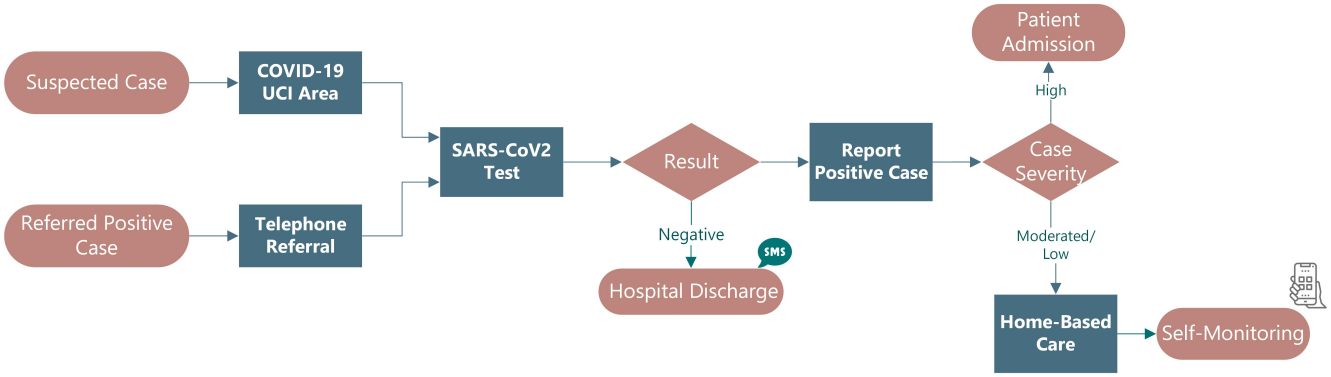

Figure 1. Flow of the referenced or suspected COVID-19 patient.

In this context, the main motivation for this article focuses on exploring and exploiting the collected data in order to develop and train ML models capable of predicting COVID-19 infections, thereby supporting health professionals at health institutions. Furthermore, the CRISP-DM methodology, used for the development of Data Mining (DM) processes, was adopted to ensure that the models created are valid and replicable.

The following document is structured in five chapters: on Chapter 1, a contextualisation and framing of the work as well as the main motivation and objectives are presented; In Chapter 2, all theoretical and scientific concepts of interest for this document are presented and documented; Chapter 3 presents the CRISP-DM methodology and all steps performed; Chapter 4 presents and discusses the obtained results; Finally, Chapter 5 aims to summarise and present the main conclusions and contributions obtained through the development of this platform, as well as proposals for future work.

Interoperable Healthcare Systems and the Use of the openEHR Standard

Interoperability is the ability of two or more systems to communicate with one another without requiring extra effort from the user, sharing critical data and initiating actions on one another[8]. Any healthcare environment is made up of a variety of different types of care that are delivered by various departments and facilities. These processes are notoriously complicated and paper-based, which the HIS and technical advancements seek to address. However, as the sophistication of medical and information technology grows, so does the risk of medical errors[9]. In all exchanges and modifications over time, the range, format, value set, occurrence, and cardinality of data must be ensured.

The openEHR standard has already demonstrated its worth and adaptability in a variety of dynamically changing situations. It is based on a two-level knowledge modelling approach in which the Clinical Information Model (CIM) is built independently of the Reference Model (RM), separating clinical and technology areas and allowing for more autonomous development in each of these domains[10,11].

The CIM promotes information consistency in the clinical domain, providing structural interoperability by using archetype units as basic components. These, in turn, are used to model increasingly complicated structures, which are called to as templates. These templates are easily adaptable to a given clinical environment by reusing existing archetypes and creating new ones for the representation of concepts that were not previously modeled in this approach. Standardization and the building block approach are especially effective in hospital environments, where the involvement of experts from each of the different areas is required for data visualizations and data entry forms to be conformant to the specific setting.[1,2,12–15]. The specification of concepts through the use of terminologies and clinical guidelines, on the one hand, guarantees semantic compatibility. On the other hand, RM incorporates a set of classes that describe the generic structure of a patient's EHR, context and audit details, all versioning standards, and access to archetypes data via *locatable* class and datatype declarations to ensure syntactic interoperability[16,17].

Machine Learning to Predict Diseases

The huge volumes of data generated daily in a hospital context demand mechanisms to classify or cluster them[18]. As a result, ML has become the most widely used sub-field of AI, with techniques including reinforcement learning and deep learning. Using a range of ML methods to generate classifiers, clusters, and rules that can organize all data based on its attributes, several everyday applications use AI recommendation systems, and efforts are underway to ensure that this trend continues in healthcare.

In literature, different healthcare and nutrition projects have used ML systems with a variety of algorithms. DM studies also use ML approaches to extract knowledge. In[19], the authors have made a survey work about the diseases diagnosed by

ML techniques, such as diabetes, heart failure, hepatitis, etc. The authors have noted that Naive Bayes and SVM algorithms can be successfully implemented to predict diseases, offering the best accuracy compared to tree algorithms. Another review paper was written in[20] in the context of the human microbiome. Several works developed using ML techniques for forecasting diagnoses such as Crohn's and colorectal diseases, bacterial vaginosis, colorectal cancer, obesity, and allergies, among others, were subjected to a systematic review. In terms of user applicability, an interactive web application for diabetes prediction was developed in[21] using the Pima Indian benchmark dataset to train an Artificial Neural Network. In order to predict a diagnosis, the application records some relevant information about the users as an input to an inference system, such as glucose and blood pressure values, body mass index, age, etc.

The work published in *Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks* has implemented another DM study in a healthcare environment. Their main goal was to apply DM techniques for the prediction of heart attacks using medical profiles such as age, sex, blood pressure, and sugar levels. By using ODANB classifiers and NCC2, they were successfully able to create models that predict the risk of heart attacks[22].

Additionally, the authors of *ML in Nutritional Follow-up Research* show how a DM study was developed using a nutritional dataset and ML algorithms. The main purpose of this case study is to create a predictive model for the eventual necessity of a patient to be followed by a nutrition specialist. The CRISP-DM method was combined with data from CHUP institution. Furthermore, five ML models were tested, including Decision Trees, Support Vector Machines, Bayesian Networks, Decision Rules, and Nearest Neighbors. The researchers developed multiple models, finding the key features of a patient that predicted the need for follow-up by a nutrition specialist, thereby assisting physicians in making the optimal choices[23].

Related Work

The openEHR standard methodology has been implemented in several national healthcare systems and hospitals. One major implementation of openEHR standards is currently being deployed in the Ministry of Health in the Republic of Slovenia. The developed solution is an HIS, supporting healthcare systems from LSs, with the ability to transform data into the openEHR format. As of this document's publication, more than 85% of Slovenia's national heath data is saved on the developed platform[24]. Additionally, Wales's National Health Service (NHS) has been carrying out a technical evaluation of openEHR's ability as a repository for structured clinical data, aiming to roll it out to support national projects and provide shared medication records for NHS Wales[25].

The Obscare platform, a Portuguese software, uses OpenEHR's ability to represent an obstetric-specific EHR and its ability to represent clinical concepts. Their analysis shows that openEHR's CKM repository still needs further work to be able to fully answer obstetrics' needs. There are still obstetric archetypes to be modelled, and edits may be required for those that already exist. Afef S. Ellouzea, Sandra H. Tlilia, and Rafik Bouazizb developed a new methodology for generating interfaces based on openEHR archetypes called *OpenEHR modeling Methodology (OpenEHR-MM)*[26].

Regarding IoT devices, the writers of *Open IoT architecture for continuous patient monitoring in emergency wards* have proposed an open architecture to track the physiological parameters of patients, using open protocols from the wearable sensors up to the monitoring system. In the IoT device aspect, the authors rely on the oneM2M technical standard for interoperability regarding architecture, API specifications, and security for M2M/IoT technologies, while employing openEHR for data semantics, storage, and making health data available to medical personnel in their EHR[27].

Tarenskeen, Debbie and van de Wetering, Rogier and Bakker, René and Brinkkemper, Sjaak argue that Conceptual Independence (CI) contributes to flexible data models that are independent of the application side at the level of IT infrastructure flexibility.Their study was performed through the use of mixed-methods research in 10 healthcare organizations, where five of them have implemented openEHR. All the studies converge to the same conclusions when demonstrating that the systems based on openEHR have greater capacity for change and remodelling. In addition, these organizations have shown a positive effect on the reuse of functionality and modularity[28]. The openEHR methodology cannot be used exclusively to build a healthcare system because its implementation demands the use of IT experts to create a standardized and reliable system free of data loss.

According to approach published in *A Migration Methodology from Legacy to New Electronic Health Record Based on OpenEHR* suggests an interoperable approach through the conversion process from SQL architecture to a NoSQL scheme, while maintaining the integrity of clinical data[29].

In order to create intelligent systems, in *Automatic Conversion of Electronic Medical Record Text for OpenEHR Based on Semantic Analysis*, the authors proposed a Wide and Deep Recurrent Neural Network (WDRNN) algorithm to automatically convert free text into structured electronic records, based on the ML classification approach. Besides, the authors also presented a CRF-RNN Label Model (CRLM) to improve the accuracy of entities and clinical concepts[30].

Methods

SARS-CoV2, the newest coronavirus, caused an unexpected global epidemic in late 2019 and early 2020. As the pandemic spread, each country's healthcare system had to adapt fast to new cases. As a contingency plan, COVID-19 screening questions based on clinical symptoms were included. This study proposes an openEHR-based system for tracking symptoms and health conditions of suspected or confirmed SARS-CoV-2 infected patients. This system was implemented in the CHUP institution, and patients could use it to track their clinical status at home either before or after learning their SARS-CoV-2 test result. The main focus of the present study is to extract useful patterns and knowledge from the data of inquired patients in the hopes of developing a predictive model capable of distinguishing between healthy people and people with COVID-19, thus assisting healthcare professionals in the early detection of infected patients, allowing them to isolate as soon as possible and consequently decreasing the spread of the virus. To perform a more detailed analysis and extract new knowledge about Portugal's epidemiological situation from the data generated by the novel system, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was determined to be appropriate for its processing and analysis.

Ethics

Between March 2020 and January 2021, all patients who submitted data from the dataset under study gave their consent for their data to be processed in accordance with the provisions of article 9, paragraph 2, c) of the Portuguese General Data Protection Regulation (GDPR) and article 29 of law 58/2019 (Portuguese national law implementing the GDPR), as well as in their consent for the use of the *CHUPCovid* application, article 9, paragraph 2, a) of the GDPR of the Portuguese legislation[31].

The authors were responsible for the data acquisition and processing software, according to CHUP hospital requirements. The study was reviewed by the interdisciplinary Hospital's Information Technology Committee, in Portuguese *Comissão de Informática*, which checked data and procedural conformity with current ethical and legal guidelines. Data anonymization was also used in this investigation, assuring security and transforming personal data into anonymous data. The data collection under study was processed to remove and change information that could identify a person. This method produced entirely anonymised data that cannot be linked to any individual. Furthermore, the authors have an authorization signed by the president of the CHUP institution to use the dataset and publish the study's results. A physician and a nurse, both authors of the current paper, also guaranteed that the hospital's ethical and data protection rules were followed.

CRISP-DM

CRISP-DM is a popular framework used for developing DM processes worldwide. This methodology was financed by the European Community and allows for project replication as well as project planning and management[32]. CRISP-DM is a cyclical process comprised of six stages, as shown in Figure 2: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment[33]. The next subsections contain a detailed description of each of these phases, with the exception of the last phase, the deployment of the model.

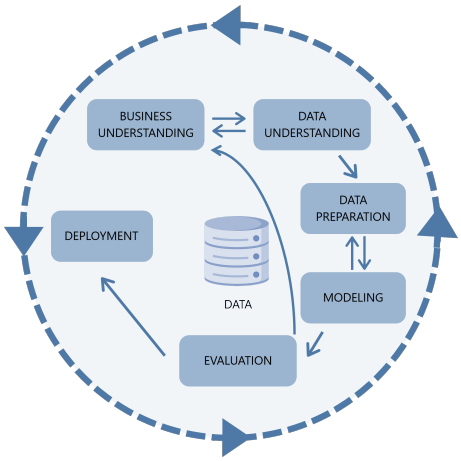

Figure 2. Stages of the CRISP-DM methodology. Adapted from[33].

Business Understanding

COVID-19 symptoms include fever, headache, respiratory symptoms such cough and dyspnea, and loss of taste or smell. Because it's a new disease with variable manifestations, diagnosing it is challenging because most of the symptoms are moderate, common daily conditions for some people, such as headaches, or symptoms of common diseases, such as the flu.

Just because of that, COVID-19 is diagnosed by laboratory tests. Thus, a model that can detect COVID-19 using a patient's clinical symptoms could greatly improve the health system's capacity to act quickly and efficiently to new cases. The purpose of this study is to investigate which factors influence COVID-19 diagnosis and to develop a predictive model for early disease detection using clinical symptomatology data from patients. To build models that can extract relevant information from patient data, this study will apply ML algorithms. The goal is to improve patient care and get them the right treatment as quickly as feasible.

Data Understanding

In order to fully understand the data and discover relationships between attributes, it was essential to go through this stage. The dataset used in this study contains a range of information extracted from an openEHR-based system for tracking symptoms and health conditions of suspected or confirmed SARS-CoV-2 infected patients that were being treated at the CHUP. The data in the anonymised dataset under study only refers to patient submissions made between March 2020 and January 2021, for a total of 13,434 instances and 14 attributes. Each instance corresponds to an inquired-about patient and contains his/her medical data. The dataset under study is composed of 4 integer attributes, 9 polynomial attributes, and 1 binomial attribute that correspond to the COVID-19 test results, which are described in the table 1.

Attribute	Description	Type
Patient_id	Patient's Identifier	Integer
Age	Patient's age	Integer
Gender	Patient's gender [1]	Integer
Temperature	Patient's body temperature [2]	Integer
Headache	Patient's headache evaluation [2]	Polynomial
Muscle_pain	Patient's muscle pain evaluation [2]	Polynomial
Cough	Patient's cough evaluation [2]	Polynomial
Diarrhea	Patient's diarrhea evaluation [2]	Polynomial
Thoracalgia	Patient's thoracalgia evaluation [2]	Polynomial
Shortness_of_breath	Patient's shortness of breath evaluation [2]	Polynomial
Shortness_of_smell_taste	Patient's shortness of smell and taste evaluation [2]	Polynomial
Medication_last_24h	Medications taken in the previous 24 hours [2]	Polynomial
Global_evaluation	Patient's health status [2]	Polynomial
Result	COVID-19 test result [3]	Binomial

[1] {Female, Male} [2] {No,I have now, Keeps, Improved, Worsened} [3] {Negative, Positive}

Table 1. Description of the attributes of the dataset under study

According to age, gender, and temperature, the data in the dataset under investigation is represented in the graphs in the following figures in a graphical format. The figure 3 depicts the age distribution of people starting from newborns to the elderly, with the most records occurring between the ages of 25 and 60 years, as seen below.

Figure 3. Distribution of patients per age.

The fact that the COVID-19 screening questionnaires were available online suggests that the increased concentration of information in this age group may be a result of the younger generation's greater familiarity with technological advances.

In terms of patient's gender, Figure 4 shows that the female gender was more prevalent than the male gender in the research dataset.

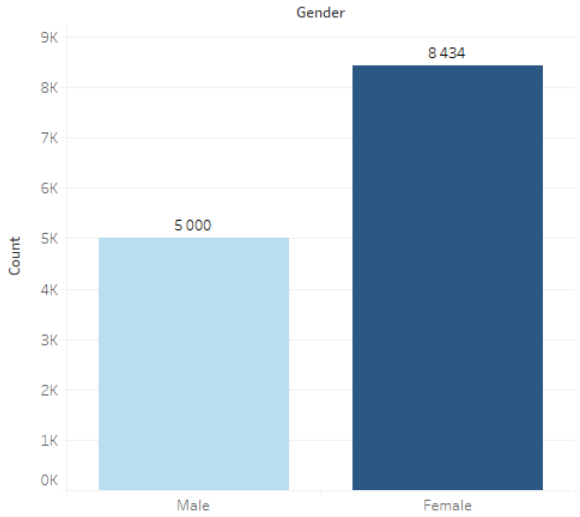

Figure 4. Distribution of patients per gender.

One of the most important established metrics is the patient's body temperature, which was calculated by the patient himself/herself. Figure 5 describes the temperature distribution of the patients, ranging from 35 to 42 degrees Celsius. The analysis of this figure reveals that the majority of patients have a temperature of 36 degrees Celsius, and by plotting a perpendicular axis at that point, a nearly symmetrical distribution of temperatures to either side can be seen. Unexpectedly, most patients have a normal body temperature, concentrating between 35 and 37 degrees Celsius. This is surprising given that the majority of the individuals considered in this study are COVID-19 positive cases, as it will be seen below.

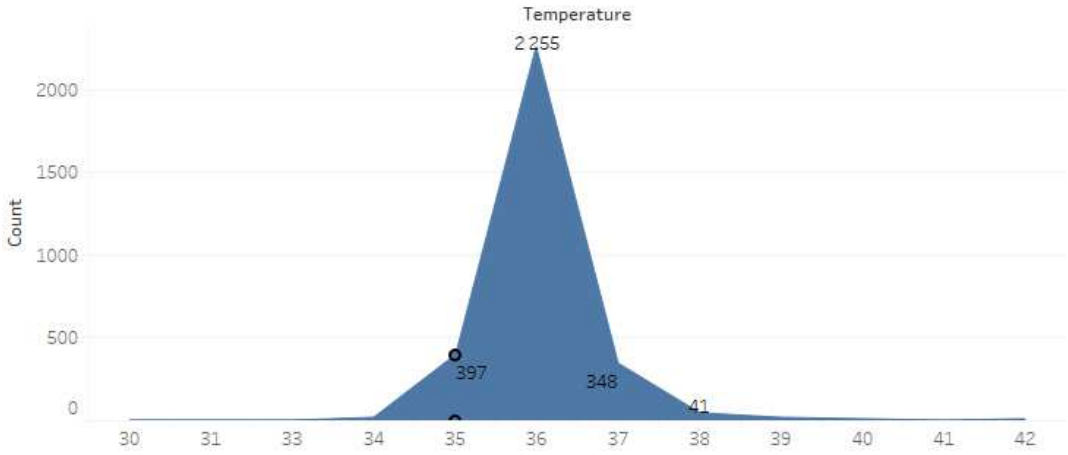

Figure 5. Distribution of patients per body temperature.

Finally, Figure 6 represents the distribution of patients of this case study in terms of COVID-19 results (the target attribute). This figure shows that the target class distribution is highly imbalanced, with only 7.24% of occurrences yielding a negative result and the remaining 92.76% yielding COVID-19 positive cases.

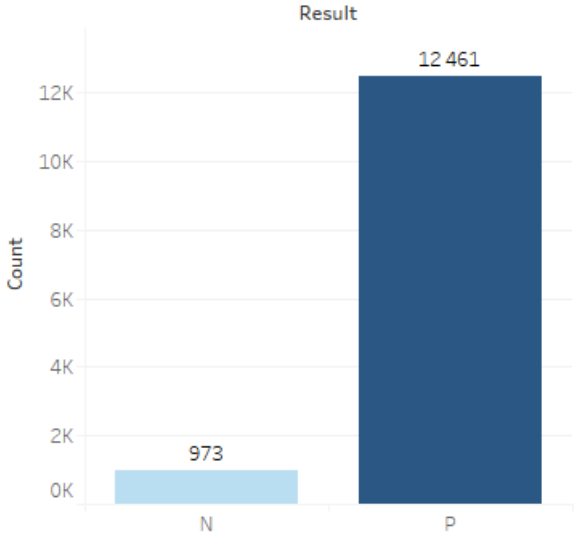

Figure 6. Distribution of patients per test result (target).

These numbers reveal that the majority of users who responded to the screening questionnaire provided in the developed web application only wanted to be remotely monitored in case they tested positive for COVID-19. As a result, an unbalanced dataset was produced.

Data Preparation

As the longest phase of the entire CRISP-DM process, data preparation concerns the integration, cleaning, transformation, and sampling of data. The data was initially incorporated and the data cleaning process was applied, analyzing the existence of duplicate data, missing values, outliers, and inconsistencies.

During the data cleaning process, no duplicate data or outliers were discovered. However, some inconsistencies were identified that had to be addressed.Most inconsistencies were found in the *Temperature* attribute. Because this is a numerical attribute, it is prone to some disparity, such as some patients filling in the values with commas and others rounding them, as well as some putting the unit of measurement and others don't, and in the case of putting the unit, the formatting may differ, i.e. "ºC", "degrees", and "degrees Celsius". As a result, all units were removed and all temperature values were rounded in order to convert them to the Integer type. In addition to this attribute, the attribute *Medication_last_24h* also required a specific transformation process because, since it is a free text field, several designations were used by the patients to designate the same medication. As a result, a lengthy and laborious transformation process was undertaken to ensure that the drug names were consistent.

In addition to inconsistencies in the data, some missing values were discovered, which were treated by replacing them with the mean for numeric attributes and the mode for nominal attributes. Hence, the *Temperature* missing values were replaced by the average value of this attribute. In turn, the missing values of nominal attributes that corresponded to the patient's symptoms were replaced by the most frequent value, which was 'No'.

The under and over types of sampling methods were evaluated in this phase for the definition of different data approaches in order to investigate which type of sampling is better for the classification of COVID-19 cases. In the next stage, Modeling, different scenarios will be generated by selecting certain attributes in order to investigate their impact on the final prediction.

Modeling

This phase consisted in the preparation of different DM Models (DMM) using the *RapidMiner* with the dataset resulting from the Data Preparation stage. Each DMM can be described as belonging to an Approach (A), being composed by a Scenario (S), a Missing Values Approach (MVA), a DM Technique (DMT), a Sampling Method (SM), a Data Approach (DA) and a Target (T), as expressed in Eq. 1.

$$DMM = \{A, S, MVA, DMT, SM, DA, T\} \tag{1}$$

There was only one target (T), which was the *result* variable. Since Classification was the chosen Approach (A), six different classifiers were selected to be used as DMTs, namely Decision Tree (DT), Random Forest (RF), Random Tree (RT), Naive

Bayes (NB), Naive Bayes - Kernel (NB-K) and Deep Learning (DL). The DL algorithm is an implementation of the Rapidminer operator which is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation on a node of H2O cluster.

For each DMT, three Sampling Methods (SM) were tested:

- *Split Validation*, with 80% of the data used for training and the remaining amount for testing.
- *Split Validation*, with 70% of the data used for training and the remaining amount for testing.
- *Cross Validation*, using 10 folds and where all data is used for testing.

Because the target variable's class distribution was considerably unbalanced, two Data Approaches (DA) were investigated: undersampling and oversampling, with the SMOTE upsampling methodology being used.

Regarding the scenarios, the first scenario (S1) includes all attributes. In the second scenario (S2), it was decided to remove the *Thoracalgia* attribute. On the other hand, the third scenario (S3) includes all attributes except the *Shortness_of_smell_taste* attribute since the loss of smell does not always imply the loss of taste, and vice versa. Therefore, it was decided that it was important to investigate the influence of this attribute on the prediction process.

As a result, in this study, the DMMs are defined as follows:

- A = {Classification}
- S = {S1, S2, S3}
- MVA = {N/A, Replace (Average and Replenishment)}
- DMT = {DT, RF, RT, NB, NB-K,DL}
- SM = {Split Validation (80%), Split Validation (70%), Cross Validation (10 folds)}
- DA = {Undersampling, Oversampling (SMOTE upsampling)}
- T = {result}

In total, 216 models were induced according to Eq. 2.

$$DMM = 1(A) \times 3(S) \times 2(MVA) \times 6(DMT) \times 3(SM) \times 2(DA) \times 1(T) \tag{2}$$

Evaluation

With the classification approach, each model generated a confusion matrix for evaluation, which represents the number of False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN) results for the model being evaluated. It is possible to calculate a variety of evaluation metrics from these values, however this study, in particular, used the accuracy (3) and precision (4) metrics, as well as the sensitivity (5) and specificity (6) metrics, to support the evaluation and conclusion of the research case. Each one of these measures is described in detail below, along with how they are calculated.

– **Accuracy:** This indicator calculates the ratio between the instances correctly classified by the forcast model and all classified instances for the correctly TP classified instances, that responds to the question:

How many patients were accurately classified out of the total?

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

– **Precision:** This parameter measures the proportion of positive occurrences properly classified by the model to the total number of positive instances, that responds to the question:

How many of patients who were classified with COVID-19 actually had the disease?

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

– **Sensitivity:** This metric is considered as an integrator indicator and measures the ratio of positive instances correctly classified by the model to the total positive instances, that responds to the question:

How many COVID-19 patients were successfully predicted out of all of them?

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

– **Specificity:** Reveals the correctly TN classified instances through the calculation of the proportion of negative occurrences correctly classified by the model to the total negative instances, that responds to the question:

How many healthy patients were correctly predicted?

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

Results

The results analysis is subdivided by the different metrics to be evaluated, which were previously described in the evaluation stage of the CRISP-DM cycle. Each DMT's best outcome for each metric was analyzed after six predictive algorithms were tested. In terms of Accuracy, i.e., the assertiveness of patient labeling, Table 2 shows that S1 was the one with the best results, which combined with DT, RF, NV-K and NV DM techniques achieved results above 85%. It is worth mentioning that the S3 scenario, the one with the removal of the *Shortness_of_smell_taste* attribute, combined with the DL algorithm also obtained a good result with 89,07% of accuracy. Finally, it is important to note that the best results are not associated to a MVA.

DMM	DMT	S	SM	MVA	DA	Accuracy (%)
3	DT	S1	Split Validation (80%)	N/A	SMOTE	96,25
7	RF	S1	Split Validation (70%)	N/A	SMOTE	91,36
105	DL	S3	Split Validation (80%)	N/A	SMOTE	89,07
27	NV-K	S1	Split Validation (80%)	N/A	SMOTE	87,45
19	NV	S1	Split Validation (70%)	N/A	SMOTE	86,32
13	RT	S1	Split Validation (70%)	N/A	SMOTE	68,26

Table 2. DMMs with the highest accuracy for each DMT.

Table 3 shows that the precision metric obtained excellent results, with all models scoring over 90%, indicating that the patients identified by COVID-19 have the condition. The DMM3, using the DT algorithm, the S1 scenario, Split Validation (80%), and SMOTE Upsampling, had the best Precision result at 99.91%.

DMM	DMT	S	SM	MVA	DA	Precision (%)
3	DT	S1	Split Validation (80%)	N/A	SMOTE	99,91
13	RT	S1	Split Validation (70%)	N/A	SMOTE	98,99
21	NV	S1	Split Validation (80%)	N/A	SMOTE	98,80
63	NV-K	S2	Split Validation (80%)	N/A	SMOTE	98,36
103	DL	S3	Split Validation (70%)	N/A	SMOTE	97,57
83	RF	S3	Cross Validation	N/A	SMOTE	91,97

Table 3. DMMs with the highest precision for each DMT.

In contrast, the worst model, DMM83, which employs the RF algorithm, the S3 scenario, the SMOTE data approach and the Cross Validation SM, achieved a Precision of 91,97%, which is still quite high.

Regarding the sensitivity measure, which results are presented in Table 4, the predictive models with the best performance were RF, DT, DL and NV-K. This means that these models ensure that at least 82% of patients infected with COVID-19 were

DMM	DMT	S	SM	MVA	DA	Sensitivity (%)
9	RF	S1	Split Validation (80%)	N/A	SMOTE	93,42
3	DT	S1	Split Validation (80%)	N/A	SMOTE	92,58
105	DL	S3	Split Validation (80%)	N/A	SMOTE	89,37
28	NV-K	S1	Split Validation (80%)	Replace	SMOTE	82,57
22	NV	S1	Split Validation (80%)	Replace	SMOTE	79,09
18	RT	S1	Cross Validation	Replace	SMOTE	50,80

Table 4. DMMs with the highest sensitivity for each DMT.

successfully predicted. Sensitivity is the most important metric to evaluate models in the scope of this study, where it is harmful to predict that a patient infected with SARS-CoV-2 is healthy. It is worth noting that this is the only metric in which DMM3 did not produce the best results, but its performance was still adequate, being the second DMM with the highest sensitivity value - 92,58%.

On the other hand, Specificity indicates how many healthy patients were appropriately predicted. Table 5 contains the DMMs with the best specifity results. The DMM3, the one characterized by the DT algorithm, the S1 scenario, the Split Validation with 80% of the data used for training and the SMOTE Upsampling technique, reveals the best combination to achieve the highest Specificity result - 99,92%.

DMM	DMT	S	SM	MVA	DA	Specificity (%)
3	DT	S1	Split Validation (80%)	N/A	SMOTE	99,92
13	RT	S1	Split Validation (70%)	N/A	SMOTE	99,63
21	NV	S1	Split Validation (80%)	N/A	SMOTE	99,12
63	NV-K	S2	Split Validation (80%)	N/A	SMOTE	98,76
103	DL	S3	Split Validation (70%)	N/A	SMOTE	98,18
83	RF	S3	Cross Validation	N/A	SMOTE	92,35

Table 5. DMMs with the highest specificity for each DMT.

Discussion

According to the majority of the study's rusults, Split Validation was the most successful SM in this research. When dealing with huge datasets and complex preparation processes, split validation is advantageous since it allows for some uncertainty about the model's robustness to be accepted. Although it requires more fully tested models than other methods, Cross Validation is more computationally complex when the data set is huge. This results in a slower overall computational performance.

In terms of scenarios, it can be observed that scenario S1, which featured all of the attributes, generated the best results, followed by scenario S3, which made it into six of the best performing models and scenario S2, which made it into two of the best performing models. In other words, while evaluating the diagnosis of COVID-19 cases, all of the attributes that were used in this study should be taken into account.

The MVA can be seen that perform better in general when the missing values are not replaced by the mean or mode value, depending on whether the missing values are numerical or nominal, as can be seen in the example above. Because of the large number of missing values in the dataset, which contains a large number of categorical attributes corresponding to the progress of symptoms that are extremely subjective in nature, this is not surprising.

The Oversampling approach, along with the SMOTE Upsampling technique, produced the greatest results by far. There has been some testing with the Undersampling method, but it did not generate the best results. This was most probably caused by the fact that the minority class was extremely small in relation to the majority class, resulting in a considerable loss of critical data.

Regarding algorithms, the DT was unquestionably the most accurate. The performance of the remaining methods cannot be easily compared because of the complexity of the problem. The average of all models trained by each forecasting algorithm and sampling strategy was used to produce dashboards for each indicator, which made it easier to evaluate the results in general.

As seen in figure 7, RF with Split Validation (70%) and Deep Learning with Split Validation (80%) were the two DMMs that obtained better average accuracy. In terms of Precision, the DMM using the RT algorithm and the Split Validation (70%) method achieved the highest result - 98,51%, followed by the DMM using the NB classifier and the Split Validation (80%) method, which reached a 91,92% value.

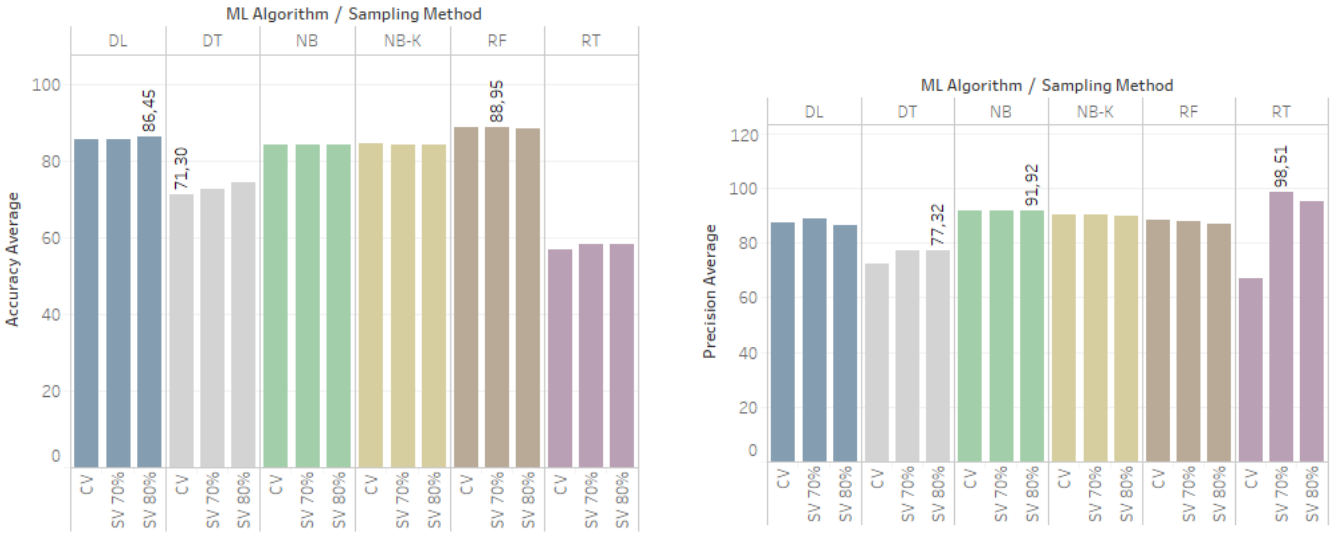

Figure 7. Average accuracy and precision of 216 tested models per DMT and SM.

Similarly to the accuracy metric, figure 8 shows that Deep Learning (86,47%) and Random Forest (90,58%) stand out among the other tested algorithms in terms of Sensitivity. On the other hand, the RT algorithm's low results, all below 50%, are worth emphasizing.

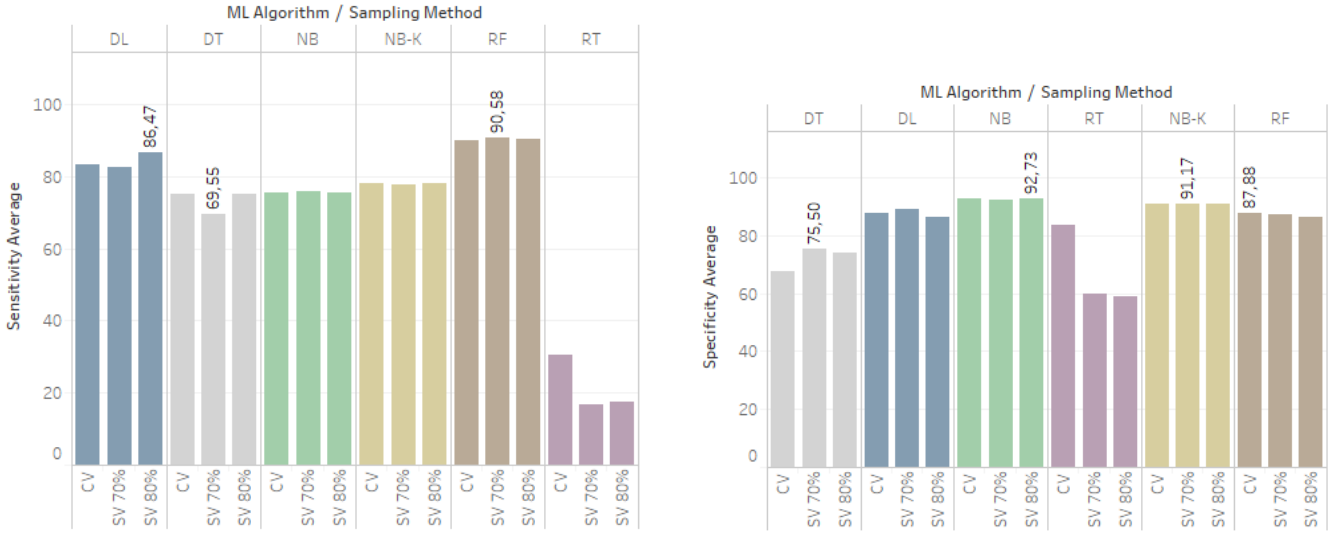

Figure 8. Average Sensitivity and Specificity of 216 tested models per DMT and SM.

For the specificity metric, as shown in figure 8, the NB-K and NB algorithms generated the best results overall, with values ranging between 91 and 93% of the total possible outcomes. The RT algorithm, on the other hand, provided averages that were extremely low. As mentioned earlier, when using the sampling method Split validation instead of Cross validation, the average of the indicators also produces better results.

Conclusions

IT systems are changing the healthcare industry in ways never thought before, from the discovery of cures for diseases and the development of new treatment techniques to the improvement of patients' diagnoses and their enhancement. As a consequence, the benefits of using IT approaches in clinical procedures, such as improving the patients' quality of care and optimising the health institution's resources, have become widely recognized, from health centres to large-scale hospitals around the world.

Accordingly, one of the most promising outcomes of this project is the discovery of how quickly and efficiently globally recognized methodologies and standards such as openEHR can be implemented, as well as how they can interoperate with LSs already in place at each health institution.

Through the methodologies and investigation strategies chosen, it was possible to delineate a valid strategy starting from topics and key ideas that became more solid and justified with the revision of the literature. Additionally, this study demonstrated that the data generated by this new system can be used to train predictive ML models with acceptable performance. In this context, the openEHR standard was quickly adapted and implemented in a COVID-19 patient circuit, and the data from inquired patients was used to feed forecasting models based on the symptoms and current health state of the patients.

In terms of results, practically all models achieved accuracy rates of over 80%, which is remarkably impressive. DMM number 3 had the best results with 96.25% accuracy, 99.91% precision, 92.58% sensitivity, and 99.92% specificity, combining the dataset with all symptoms, the Split Validation (80%) method, and the Decision Tree algorithm without replacing the missing values. Hence, after collecting more data and subjecting the models to additional testing and rigorous evaluations, the predictive model could be later implemented in a CDSS to assist healthcare professionals.

For future work, it is suggested to change some occurrences in the items of the openEHR templates developed, thus implying the need to fill them in order to reduce the number of missing values in the dataset, and therefore improve the trustworthiness of the produced results.

Acknowledgements

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

Author contributions

D.O., D.F., and N.A. wrote the main manuscript text and created all of the figures and tables. All of the authors reviewed the manuscript. A formal analysis was conducted by P.L. and A.A., and the final draft was supervised and revised by J.M..

Competing interests

The authors declare no competing interests.

References

1. Fetter, M. S. Interoperability—making information systems work together. *Issues mental health nursing* **30**, 470–472 (2009).
2. Esteves, M., Esteves, M., Abelha, A. & Machado, J. A proof of concept of a mobile health application to support professionals in a portuguese nursing home. *Sensors* **19**, 3951 (2019).
3. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *Jama* **309**, 1351–1352 (2013).
4. Lee, C. H. & Yoon, H.-J. Medical big data: promise and challenges. *Kidney research clinical practice* **36**, 3 (2017).
5. Luo, J., Wu, M., Gopukumar, D. & Zhao, Y. Big data application in biomedical research and health care: a literature review. *Biomed. informatics insights* **8**, BII–S31559 (2016).
6. Oliveira, D. *et al.* Openehr modeling: improving clinical records during the covid-19 pandemic. *Heal. Technol.* 1–10 (2021).
7. Oliveira, D. *et al.* Management of a pandemic based on an openehr approach. *Procedia Comput. Sci.* **177**, 522–527 (2020).
8. Cardoso, L. *et al.* The next generation of interoperability agents in healthcare. *Int. journal environmental research public health* **11**, 5349–71, DOI: 10.3390/ijerph110505349 (2014).
9. Miranda, M., Duarte, J., Abelha, A. & Machado, J. Interoperability and Healthcare. *Eur. Simul. Model. Conf. 2009* 205–212 (2009).

10. Pedersen, R., Granja, C. & Marco-Ruiz, L. Implementation of openehr in combination with clinical terminologies: Experiences from norway. *Int J Adv Life Sci* **9**, 82–91 (2017).

11. de Moraes, J. L. C., de Souza, W. L., Pires, L. F. & do Prado, A. F. A methodology based on openehr archetypes and software agents for developing e-health applications reusing legacy systems. *Comput. methods programs biomedicine* **134**, 267–287 (2016).

12. Tute, E., Wulff, A., Marschollek, M. & Gietzelt, M. Clinical information model based data quality checks: Theory and example. In *EFMI-STC*, 80–84 (2019).

13. Yang, L., Huang, X. & Li, J. Discovering clinical information models online to promote interoperability of electronic health records: A feasibility study of openehr. *J Med Internet Res* **21**, e13504, DOI: 10.2196/13504 (2019).

14. Sahakian, T. *et al.* | the fine line between decisions and evidence-based decisions: Contextualizing and unraveling the evidence-based management process in hospital settings. *Evidence-Based Manag. Hosp. Settings* 74.

15. Rawat, R. & Yadav, R. Big data: Big data analysis, issues and challenges and technologies. In *IOP Conference Series: Materials Science and Engineering*, vol. 1022, 012014 (IOP Publishing, 2021).

16. OpenEHR. openEHR Specification Components.

17. Hak, F. *et al.* An openehr adoption in a portuguese healthcare facility. *Procedia Comput. Sci.* **170**, 1047 – 1052, DOI: https://doi.org/10.1016/j.procs.2020.03.075 (2020). The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

18. Neves, J. *et al.* A Deep-Big Data Approach to Health Care in the AI Age. *Mob. Networks Appl.* **23**, 1123–1128, DOI: 10.1007/s11036-018-1071-6 (2018).

19. Fatima, M. & Pasha, M. Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* **9**, 1 (2017).

20. Marcos-Zambrano, L. J. *et al.* Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. microbiology* **12**, 313 (2021).

21. Dey, S. K., Hossain, A. & Rahman, M. M. Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In *2018 21st international conference of computer and information technology (ICCIT)*, 1–5 (IEEE, 2018).

22. Srinivas, K., Rani, B. & Govrdhan, A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *Int. J. on Comput. Sci. Eng.* **02**, 250–255, DOI: 10.1.1.163.4924 (2010).

23. Reis, R., Peixoto, H., Machado, J. & Abelha, A. Machine Learning in Nutritional Follow-up Research. *Open Comput. Sci.* **7**, 41–45, DOI: 10.1515/comp-2017-0008 (2017).

24. Better. Client Stories - Ministry of Health of Republic of Slovenia.

25. Meredith, J. What is openEHR and why is it important? | Digital Health Wales (2021).

26. Alves, D. S. *et al.* Can openEHR represent the clinical concepts of an obstetric-specific EHR - Obscare software? *Stud. Heal. Technol. Informatics* **264**, 773–777, DOI: 10.3233/SHTI190328 (2019).

27. Pereira, C. *et al.* Open IoT architecture for continuous patient monitoring in emergency wards. *Electron. (Switzerland)* **8**, 1–15, DOI: 10.3390/electronics8101074 (2019).

28. Tarenskeen, D., van de Wetering, R., Bakker, R. & Brinkkemper, S. The contribution of conceptual independence to it infrastructure flexibility: the case of openehr. *Heal. Policy Technol.* **9**, 235–246 (2020).

29. Khennou, F., Chaoui, N. E. H. & Khamlichi, Y. I. A migration methodology from legacy to new electronic health record based openehr. *Int. J. E-Health Med. Commun. (IJEHMC)* **10**, 55–75 (2019).

30. Zhu, Y., Jin, X. & Li, L. Automatic conversion of electronic medical record text for openehr based on semantic analysis. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, 35–39 (IEEE, 2019).

31. Lei 58/2019, 2019-08-08 - dre. https://dre.pt/pesquisa/-/search/123815982/details/maximized. (Accessed on 09/30/2021).

32. Martins, B., Ferreira, D., Neto, C., Abelha, A. & Machado, J. Data mining for cardiovascular disease prediction. *J. Med. Syst.* **45**, 1–8 (2021).

33. Ferreira, D., Silva, S., Abelha, A. & Machado, J. Recommendation system using autoencoders. *Appl. Sci.* **10**, 5510 (2020).