machine learning & knowledge extraction

Article

Extracting Interpretable Knowledge from the Remote Monitoring of COVID-19 Patients

Melina Tziomaka [1], Athanasios Kallipolitis [1], Andreas Menychtas [1], Parisis Gallos [2], Christos Panagopoulos [2], Alice Georgia Vassiliou [3], Edison Jahaj [3], Ioanna Dimopoulou [3], Anastasia Kotanidou [3] and Ilias Maglogiannis [1,*]

[1] Department of Digital Systems, University of Piraeus, 185 34 Piraeus, Greece; tziomakamel@unipi.gr (M.T.); nasskall@unipi.gr (A.K.); amenychtas@unipi.gr (A.M.)
[2] BioAssist SA, 265 04 Rio, Greece; parisgallos@bioassist.gr (P.G.); cpan@bioassist.gr (C.P.)
[3] 1st Department of Critical Care Medicine and Pulmonary Services, School of Medicine, National and Kapodistrian University of Athens, Evangelismos Hospital, 106 76 Athens, Greece; alvass@med.uoa.gr (A.G.V.); ejahaj@med.uoa.gr (E.J.); idimop@med.uoa.gr (I.D.); akotanid@med.uoa.gr (A.K.)
* Correspondence: imaglo@unipi.gr; Tel.: +30-2107235521

Abstract: Apart from providing user-friendly applications that support digitized healthcare routines, the use of wearable devices has proven to increase the independence of patients in a healthcare setting. By applying machine learning techniques to real health-related data, important conclusions can be drawn for unsolved issues related to disease prognosis. In this paper, various machine learning techniques are examined and analyzed for the provision of personalized care to COVID-19 patients with mild symptoms based on individual characteristics and the comorbidities they have, while the connection between the stimuli and predictive results are utilized for the evaluation of the system's transparency. The results, jointly analyzing wearable and electronic health record data for the prediction of a daily dyspnea grade and the duration of fever, are promising in terms of evaluation metrics even in a specified stratum of patients. The interpretability scheme provides useful insight concerning factors that greatly influenced the results. Moreover, it is demonstrated that the use of wearable devices for remote monitoring through cloud platforms is feasible while providing awareness of a patient's condition, leading to the early detection of undesired changes and reduced visits for patient screening.

Keywords: SARS-CoV-2; modified medical research council dyspnea grade; machine learning; interpretability; ensemble classifiers; time-to-event analysis

check for updates

Citation: Tziomaka, M.; Kallipolitis, A.; Menychtas, A.; Gallos, P.; Panagopoulos, C.; Vassiliou, A.G.; Jahaj, E.; Dimopoulou, I.; Kotanidou, A.; Maglogiannis, I. Extracting Interpretable Knowledge from the Remote Monitoring of COVID-19 Patients. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1323–1342. https://doi.org/10.3390/make6020062

Academic Editor: Jianlong Zhou

Received: 6 March 2024
Revised: 26 April 2024
Accepted: 10 June 2024
Published: 18 June 2024

1. Introduction

A continuous increase in demands for health services, whether due to overpopulation, increasing aging, or the emergence of a pandemic, results in an ever-growing need to support health professionals in making decisions [1] and choosing treatment plans according to personal characteristics and the health status of individuals [2]. The ultimate goals of all health systems are to ensure high levels of care and contribute to quality of life [3,4]. The development of computer science and the application of new advancements in the health sector present solutions that can assist substantially in reaching these goals. Fast and effective communication infrastructures [5], small and powerful mobile devices and wearables [6], diffusion of the IoT [7], and the ability to store and handle big data [8] are some of the technologies that experienced an unexpected boom and set a prosperous ground for precision and public health medicine. Big data analysis plays an important role in the provision of user-friendly applications that support digitized healthcare routines and present information with compact visualizations [9]. This data can be, firstly, analyzed and explored with plain statistical techniques and, secondly, utilized as part of a machine learning (ML) pipeline for the discovery of important patterns. In this context, several intelligent platforms can be found in the literature that manage the collection and analysis of health-related data [10–13].

Modern technological solutions help improve the quality of life of people, and, more specifically, of vulnerable groups who face a reduction in standard of living due to health issues that concern them. In favor of improved quality of life, the use of mobile devices, wearable devices, activity trackers, and biosignal sensors has proven to increase the independence of patients with chronic diseases and/or limitations [14,15]. Moreover, the enhancement of useful personal and automatically generated data in the electronic health record of a patient provides an extended and multidimensional version of the original record [16,17].

Traditionally, analysis of clinical data has been an important tool for the science of medicine and the study of epidemiology and preventive medicine. Today, data analysis has contributed significantly to decision-making in personalized clinical interventions, treatment plans, and the development of health policies for the prevention of various diseases after also studying the characteristics of the applied populations. By applying ML techniques to real health-related data, important conclusions can be drawn for unsolved issues related to disease prognosis and diagnosis, patient risk stratification, precision medicine, and public health with high predictive accuracy [18]. Graph-based representations are often considered a significant booster of ML approaches due to their capability to retrieve information from different data entities and their interconnections [19,20]. However, it should be noted that interpretability is equally important to the efficiency of ML models. The acknowledgement that ML models have the ability to reveal to the designated stakeholders the details of their inner workings and their decision-making mechanism adds value, transparency, and validity to the predicted outcome [21]. Moreover, it is the basis upon which trust can be built for healthcare professionals (HCP) and an accelerating factor for the integration of ML systems in clinical workflow. Despite the obvious necessity for interpretability, there are still many approaches that are presented without the provision of an inherent interpretability scheme or the capability for future extension [22]. Therefore, it is highly recommended that ML systems in the healthcare domain are accompanied by an interpretability scheme that returns plausible explanations for their decisions and a straightforward connection between the cause and effect [23,24].

A major challenge of modern medicine is the personalization of healthcare based on the particularities of each individual at the levels of genes, biomarkers, response to treatment, environmental influences, personal preferences, and habits [25]. Analysis of all these combined data can contribute to the improvement of personalized decisions. In recent years, it has been observed that personalized care is quite popular with oncology patients, cardiac patients, the elderly, or people with chronic diseases. Achieving a personalized health service consumption plan requires a deep analysis of all relevant data concerning the individual and the ability to discriminate important patterns within stratified populations [26,27]. Therefore, the utilization of thoroughly curated datasets that contain the samples of a population within a subcategory can highlight more fine-grained patterns tailor-made to that specific subset.

In this work, the dataset includes COVID-19 patients with mild symptoms who are receiving remote care at home and undergoing monitoring for post-COVID-19 complications. It contains information derived from electronic health records and questionnaires conducted by HCPs. The symptom of fever is a complex physiological response initiated by the activation of certain cells of the immune system that produce cytokines [28]. During a viral infection, the host develops an immune response to contain it, and fever is one of the key diagnostic signs for screening patients potentially infected with COVID-19 [29]. Furthermore, during a viral infection, the febrile response determines the survival advantage [30]. Estimating the duration of fever can play an important role in a specialist's decision to discharge a patient from the hospital as well as in their follow-up and treatment plan.

Although the modified Medical Research Council (mMRC) Dyspnea Scale is subjective and not used in clinical decisions, the scale can be examined for its potential to reveal indications for the progression of COVID-19 disease and hence was placed at the forefront of ML data analysis. Anticipating future trends in a patient's difficulty to breathe plays an important role in the discovery of patterns for the timely admission of the patient to the clinic and the day of discharge from the hospital. A secondary goal is to provide extensive interpretability results beyond prediction, aiming to determine the factors influencing analysis of the data. This is achieved by exploiting the internal mechanisms of predictive algorithms to explore and detect the most meaningful connections between model inputs and outputs. Although the dataset is significantly imbalanced in terms of the classification task, the selected classifiers show significant robustness as demonstrated by the comparison of accuracy and balanced accuracy metrics. Moreover, interpretability results are in accordance with human expertise with reference to the most important characteristics that influence prediction. Since most of them are collected automatically, remote monitoring by cloud platforms can provide real-time measurements and awareness of the patient's condition, leading not only to the early detection and treatment of any undesired change but also to the reduction of visits to patient screening centers and hospital admissions.

The remainder of the paper is organized as follows: Section 2 describes the collected dataset and applied methods for classification and interpretability results, Section 3 validates the efficiency of the proposed methodology, Section 4 describes the use of implemented systems for data collection and analysis, and Section 5 summarizes key findings.

2. Materials and Methods

2.1. Overview of Data

To support the main use-case scenarios, a dataset of patients with mild COVID-19 symptoms was curated by the HCPs of "Evangelismos Hospital" (National and Kapodistrian University of Athens—NKUA). The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of "Evangelismos Hospital" (473/7-10-2021). The HCPs of the clinic managed the patients' enrolment in the whole procedure. They assessed patients' suitability and chose asymptomatic patients or patients with mild COVID-19 symptoms to test and evaluate telecare services during the whole duration they tested positive for SARS-CoV-2. Physicians and other HCPs assessed the data that were recorded by the patients and remotely monitored the condition of mild COVID-19 or asymptomatic cases. When a patient tested positive for SARS-CoV-2, a HCP was informed. After thoroughly reading and signing the informed consent form, the patient was enrolled in the study. The HCP recorded the following information:

- Age;
- Weight;
- Sex;
- Medical history, including the following:
    - ○ Comorbidities (asthma, hypertension, hyperlipidemia, diabetes, chronic obstructive pulmonary disease (COPD), and coronary heart disease);
    - ○ Smoking;
    - ○ Medications;
- COVID-19 history, including the following:
    - ○ Date of SARS-CoV-2 positive test;
    - ○ Onset of COVID-19 symptoms;
    - ○ Number of COVID-19 vaccine doses administered if vaccinated;
    - ○ Date of last COVID-19 vaccination dose;
    - ○ COVID-19 vaccine manufacturer;
    - ○ Initial symptoms;
    - ○ Previous SARS-CoV-2 infection.

Demographic and clinical features pertaining to the study population are presented in Table 1.

Table 1. Demographic and clinical features pertaining to the study population.

Demographic and Clinical Characteristics of Patients	
Variable	**Data**
Patients, N	162
Age, years (median, IQR)	51 (42–60)
Sex, N (%)	
Male	72 (44.4)
Female	90 (55.6)
Smoking status, N (%)	
Yes	87 (53.7)
No	75 (46.3)
Comorbidities, N (%)	35 (21.6)
Hypertension	33
Hyperlipidemia	33
Coronary artery disease	5
Diabetes	4
Thyroid disease	4
Asthma	4
COPD	2
Weight, kg (mean ± SD)	70.57 ± 11.94
Vaccination status, N (%)	
4 doses	1 (0.6)
3 doses	113 (69.8)
2 doses	37 (22.8)
1 dose	5 (3.1)
Unvaccinated	6 (3.7)
Vaccine type, N (%)	
Pfizer	142 (87.6)
Moderna	5 (3.1)
Johnson & Johnson	5 (3.1)
AstraZeneca	4 (2.5)
Days since the last vaccine dose (median, IQR)	120 (88–160)
Previous infection, N (%)	6 (3.7)
Days from positive test prior to enrolment (median, IQR)	3 (1–11)
Days of symptoms prior to enrolment (median, IQR)	3 (2–5)
Days of monitoring (median, IQR)	14 (13–15)

The timeline for COVID-19 patients' data collection was decided to be 14 days, which was deemed appropriate since the hospitalization duration has been reported to be 8–16 days. The patients performed biosignal measurements (vital signs) and self-assessments (questionnaires and symptom reporting) on a daily basis following their enrolment. The data that were daily reported by the patients were the following:

- Heart rate;
- Blood pressure;
- Oxygen saturation level;
- Body temperature;
- Respiratory rate;
- Weight;
- Glucose (if appropriate, e.g., diabetic patients).

Additionally, questionnaires containing information on the patients' daily dyspnea grade (using the mMRC Dyspnea Scale) (Table 2) [31] and symptoms were also reported as part of the patients' routine. In addition, 25% of the patients, who owned smartphones and were considered capable of using smart wearables, were equipped with activity trackers (Huawei Band 6) for continuous monitoring of their biosignals and automated ingestion of data into a data collection platform during the monitoring period. The total number of patients who participated in the study with their data is one hundred and sixty-two (162). The results of our data analysis concern the most common symptoms presented in COVID-19 mild cases and their duration, as well as analytical models that can predict the daily mMRC grade, SpO_2, and fever duration.

Table 2. mMRC dyspnea scale.

mMRC Scale	
Grade	**Description**
1	No shortness of breath or shortness of breath only during strenuous work.
2	Shortness of breath when walking quickly on level ground or a slight incline.
3	You walk more slowly than people of the same age on level ground because of shortness of breath or stop for breath if you walk alone.
4	Stopping for breath after walking for about 100 m or after a few minutes of walking on a level surface.
5	Too breathless to leave the house or breathless when dressing/undressing.

2.2. The Proposed Methodology

2.2.1. Data Preparation

It is important to refer to the processes through which the raw data were submitted to properly format them for analysis. These processes include operations such as transforming their initial form as well as handling missing values. The transformation of the dataset was performed to code all symptoms from the forms and daily questionnaires through the one-hot-encoding technique. The one-hot-encoding technique generates new (binary) features, indicating the presence of each possible value from the original data to turn them into categorical variables that can be provided as input to ML algorithms. To impute the missing values, the K-Nearest Neighbor (KNN) algorithm [32] was applied with a setting of n = 5 nearest neighbors.

Regarding further data preparation and feature selection, two approaches were followed, with two different objectives, as follows:

- Classification of the mMRC Grade;
- Time-To-Event (TTE) analysis of fever remission.

For each patient, multiple questionnaires and measurements were recorded. For this reason, in the approach aimed at mMRC grade classification on a daily basis, which is derived from daily questionnaires, the dataset samples were required to correspond to daily data for each patient. Therefore, the mean value of the daily measurements corresponding to vital signs was calculated for each day and each patient, and then the data were correlated with the corresponding daily questionnaires. The final set of data exploited by the system included 1164 samples from 162 patients due to several sample deletions that did not apply to the standards. For the experiments, k models were developed, where k was set to eight to obtain a satisfying number of training samples given that the hospitalization period ranged from 8–16 days. As we proceeded in days, the number of samples with a range of 8 days later than the day that input data refer to decreases; a fact that is depicted in Table 3.

The target variable (mMRC grade) was predicted by utilizing the measurements from past days. In a series of k predictive models, where each model referred to the total architecture, presented in Figure 1, the first model utilized the measurements of the given day 1 to predict the target value of the given day. The second model used the measurements of given day 2 to predict the target value of a given day, and the same practice continued until the kth model. Attempting to predict the target value for one day away from an earlier day upon which the measurements were gathered should result in poorer performance. Each model was independent, and the generated results refer only to the specific model. As a result, at the end of the training process, k predictive models were generated, with day + k being a user-defined hyperparameter. In order to describe the health data on which this analysis is based, information about the target variable is provided in Table 3. As shown in Table 3, there is a large class imbalance for each case. The number of samples with an mMRC grade other than one or two is very few or non-existent, resulting in the binary classification task mMRC grade 1 vs. grade 2.

Table 3. Samples of target value concerning the prediction day.

Day	mMRC Grade 1	mMRC Grade 2
0	1018	146
1	938	134
2	861	120
3	791	105
4	721	92
5	647	79
6	574	68
7	520	60
8	430	48

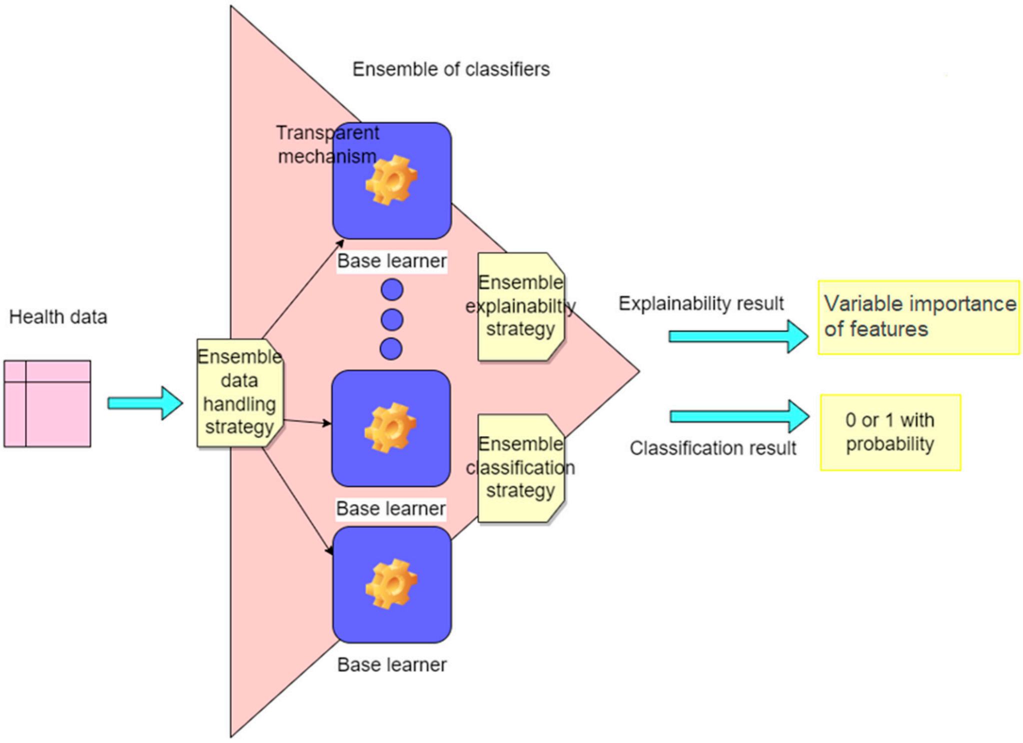


Figure 1. The workflow of the ML pipeline including classification and explanation processes.

For the TTE analysis, focused on fever remission, the data preparation was expanded to include the aggregation of daily collected patient data. This process entailed the creation of new characteristic variables that captured the cumulative days with various degrees of dyspnea and different symptoms, and the calculation of average values for biosignal measurements.

2.2.2. Classification and Interpretability Methodology

After pre-processing the dataset for classification, the variable feature values were provided to the classifier. This section describes the methodology for classifying patient data. The selection of classifiers was based on two main criteria: (a) efficiency against unbalanced datasets and (b) interpretability properties. For this reason, Extreme Gradient Boosting [33], AdaBoost [34], and Random Forest [35] classifiers were chosen, as they have demonstrated the aforementioned skills in the literature.

Extreme Gradient Boosting is a decision tree-based algorithm that has emerged in the field of applied ML due to its performance and fast execution time. It can be used for both classification and regression problems. Specifically, the method involves the sequential creation and addition of decision trees to a set, where each of them corrects the error of those that preceded it. This is a type of ensemble shape model referred to as boosting. The models are fitted using a differentiable loss function and a gradient descent optimization algorithm. The gradient of each base model's error instructs the model towards the direction and amplitude of the required modifications in order to increase the total ensemble accuracy. AdaBoost (Adaptive Boosting) is another very popular ensemble-boosting technique that aims to combine several weak classifiers to create a strong one and can be used in combination with other types of learning algorithms to improve their performance. The output of the other learning algorithms is combined into a weighted sum that represents the final result of the boosted classifier. AdaBoost has the characteristic of adaptability in the sense that weak learning models are modified in favor of those samples misclassified by previous classifiers. The individual classification models may be weak but as long as each one performs slightly better than a random guess, the final model can be shown to converge on a strong classifier. For this reason, in some problems, AdaBoost is less prone to the problem of overfitting than other learning algorithms. The Random Forest algorithm is a classification and regression method that works by building many decision trees during its training. Through the technique of an ensemble bagging scheme, it selects random observations and features to train different decision trees and then combines their decisions (mainly referred to in the literature as bagging). In classification, the output of Random Forest is the class selected from the most trees, while in regression, it is the average prediction of the individual trees. The selection of the above algorithms was made due to their performance on new data (external validation data), minimizing the generalization error. Extreme Gradient Boosting and AdaBoost minimize the deviation error, while Random Forest minimizes the variance error. Also, based on different ensemble schemes, the three classifiers consist of base learning models that provide transparency regarding their inner workings, while their combination through the ensemble scheme improves their performance and achieves state-of-the-art results. The architecture of the methodology is shown in Figure 1. The basic ensemble model workflow for generating forecasts consists of the following steps:

- Dataset preparation, where the data are partitioned into subgroups or assimilated as a whole by each underlying learning model according to the ensemble classifier strategy.
- Ensemble classification, where base classifiers are trained in parallel or serially, and their predictions are aggregated to produce the output of the ensemble classifier [36].
- Ensemble interpretability, where base classifiers return an importance value for each individual input feature in the final result, and the importance values for each feature of the base classifiers are summed following the ensemble model logic.

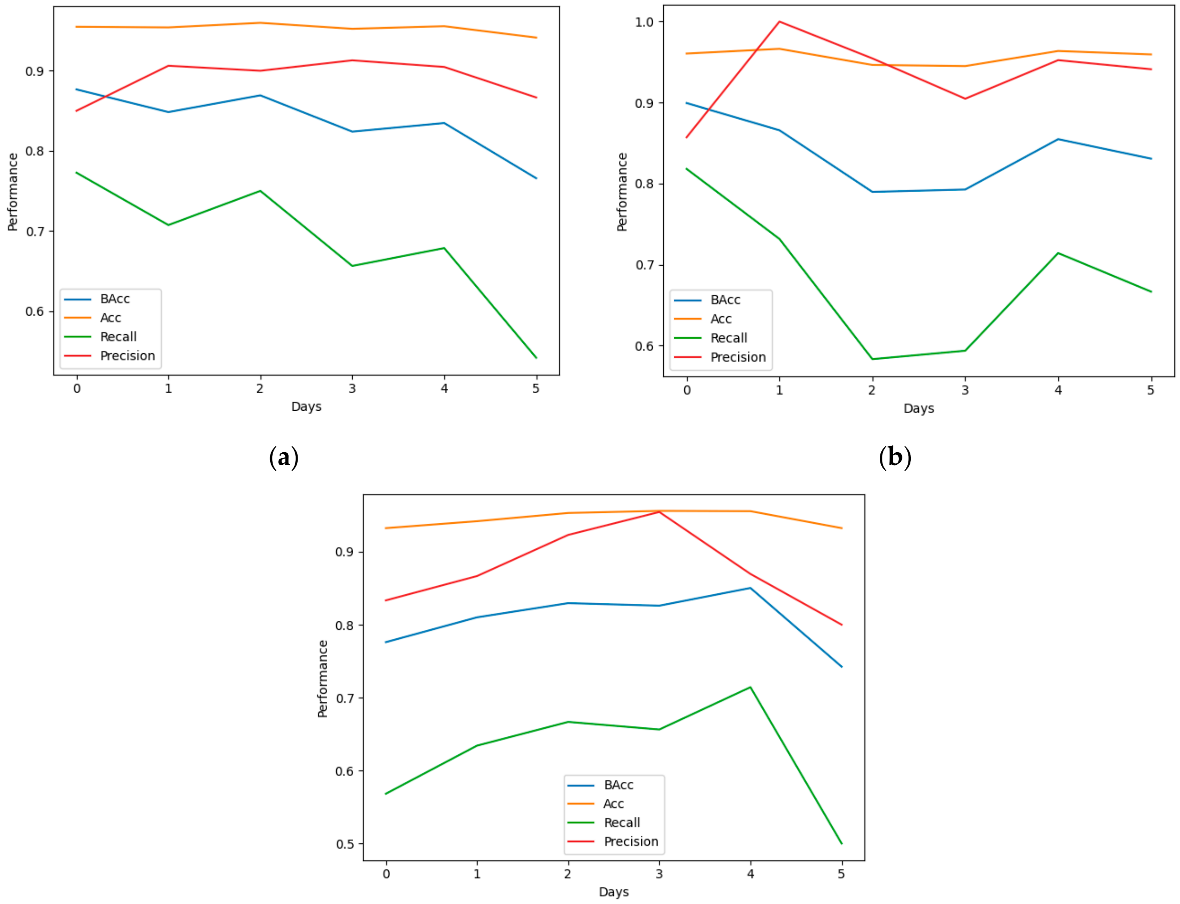
2.2.3. Time-to-Event Analysis Methodology

To analyze the expected duration of fever, the Cox regression model, also known as the proportional hazards regression model, was utilized [37,38]. The Cox proportional hazard model is a regression technique for investigating the effect of several variables on the time it takes for a particular event to occur. More specifically, the model estimates the hazard ratio of a given endpoint with a specific risk factor, which can be either a continuous or categorical variable. In each experiment, the daily questionnaires and measurements included in the training dataset were according to a specified monitoring period, and, therefore, a different prediction model was derived for each period.

3. Experimental Results

3.1. Classification

To evaluate the performance of the classification methods, the metrics of accuracy, balanced accuracy, precision, and recall were selected, and the dataset was divided into a 10-fold cross-validation scheme (10-fold cross-validation). The results of each classifier for predicting mMRC grade within a 5-day range are shown below. For each prediction of a different day, a different model was trained and tested with the corresponding inputs. This function led to the development of five prediction models. In Figure 2a–c, the value of each measurement is plotted on the y-axis, while on the x-axis the value represents the day for which the forecast was made.



(**a**) (**b**)

(**c**)

Figure 2. Performance metrics of the (**a**) Random Forest, (**b**) XGBoost, and (**c**) AdaBoost classifiers for mMRC grade classification. BAcc corresponds to balanced accuracy and Acc to accuracy.

Although the continuous line does not reflect the discrete nature of the day variable, it is used to better illustrate the trend between each day. In Figure 2a, the Random Forest classifier results are presented. The difference between accuracy and balanced accuracy metrics is expected due to class imbalance. Balanced accuracy starts at 0.895 for day 0 and reaches 0.796 for day 5. As the number of days increases, a certain decrease can be observed in all measurements. Figure 2b depicts the mMRC scoring results from the Extreme Gradient Boosting classifier. The balanced accuracy starts at 0.9 for day 0 and reaches 0.85 for day 5. Although the balanced accuracy increased for the Extreme Gradient Boosting classifier, there is a significant decrease in the recall values, which is undesirable for healthcare applications. In Figure 2c, the performance metrics of the AdaBoost classifier are presented in the same way. In this case, the recall metric has the lowest values, and there is an unexpected increase of all metrics around days 3 and 4. Balanced accuracy starts at 0.79 and reaches about 0.73 on day 5. Comparing the classification results shows the Random Forest classifier as the most efficient. Although the Extreme Gradient Boosting classifier has a slightly better balanced accuracy, Random Forest performed better on the recall metric, which is a key indicator for unbalanced medical datasets and hence the decision heavily depends on it.

3.2. Interpretability

In Figures 3–5, the graphical representations of feature variable importance for each different day for the aforementioned classifiers are presented. The variables with the highest importance score have the greatest influence in terms of the predictive outcome. A close examination of all the figures shows that some variables play little to no role in determining the outcome. These variables are (a) diabetes, (b) coronary heart disease, (c) Pfizer, (d) Moderna, (e) Johnson & Johnson, (f) AstraZeneca, and (g) "Have you ever had COVID-19". The fact that the values of these variables are of minimal importance suggests that only the selection of the remaining variables is needed to develop a simpler classification scheme. It is made evident through the comparison of the results that the importance is spread over more variables for the Random Forest classifier, while for the Extreme Gradient Boosting classifier, it is concentrated in one or two variables. For the Random Forest classifier, age, weight, body temperature, diastolic and systolic blood pressure, heart rate, and days since the last dose are the variables that most influence the classification result, while for the Extreme Gradient classifier Boosting, it is only asthma. In the case of AdaBoost, weight, age, and days since the last dose have the biggest effect. The graphical representations in Figure 6a–c are indicative of the importance of each variable in relation to days for prediction. In an attempt to distinguish variables that have a long-term daily perspective effect on the outcome, the trends of each variable are plotted in Figure 6a for the Random Forest classifier. For the Random Forest classifier, the importance is shared across multiple feature variables. The effect of systolic pressure and body temperature decreases as the number of forecast days increases, which means that a predictive model cannot base prediction for future days on these variables, while hypertension and weight are in the opposite direction. For the Extreme Gradient Boosting and AdaBoost classifiers, specific trends are not verified, as the plots are characterized by continuous fluctuations (Figure 6b,c).

3.3. Improving Models through Feature Variable Selection

As already explained in the previous subsection, the feature importance indices that the classifiers return discriminate certain features to have more influence on the classification result than others. By excluding these variables from the training process, we have the ability to generate lighter and simpler models with fewer characteristics. The excluded characteristics by means of measuring the feature importance of the Random Forest (RF) classifier (Figure 3) are the following: (a) diabetes, (b) coronary heart disease, (c) Pfizer, (d) Moderna, (f) Johnson & Johnson, (e) AstraZeneca, and (g) "Have you ever had COVID-19" (Ill before). The experiments with the RF classifier are conducted from scratch and led to the results of Figure 7. The number of selected variables was set to seven by the application of a Grid search algorithm to determine the best accuracy with a minimum number of input variables. The most important features of the new predictive models are as follows:

- Days since the last dose;
- Age;
- Asthma;
- Heart rate;
- Diastolic pressure;
- Systolic pressure;

- Body temperature;
- Weight.

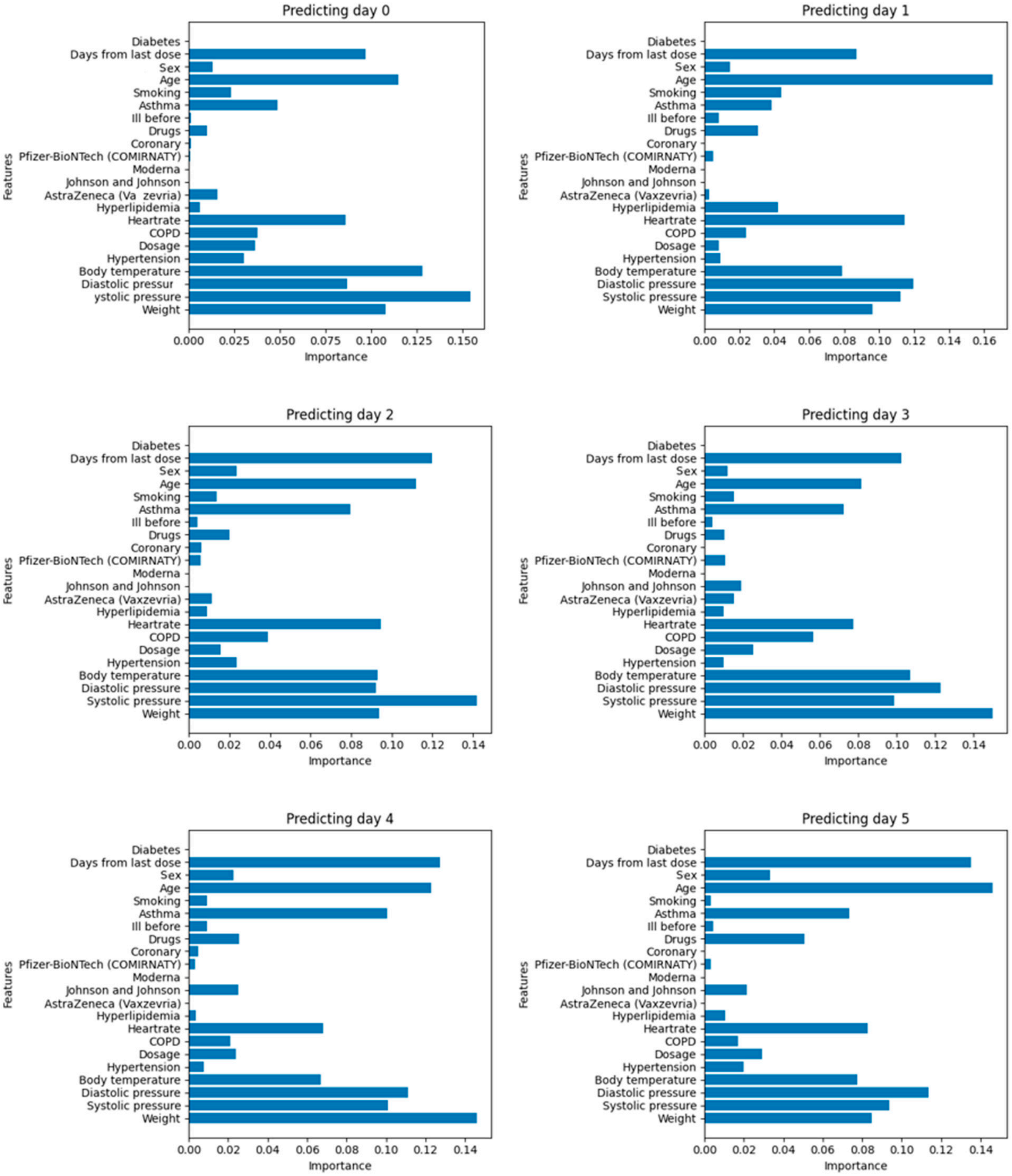


Figure 3. The importance score of variable features of the Random Forest classifier for predicting the mMRC grade in the coming days.

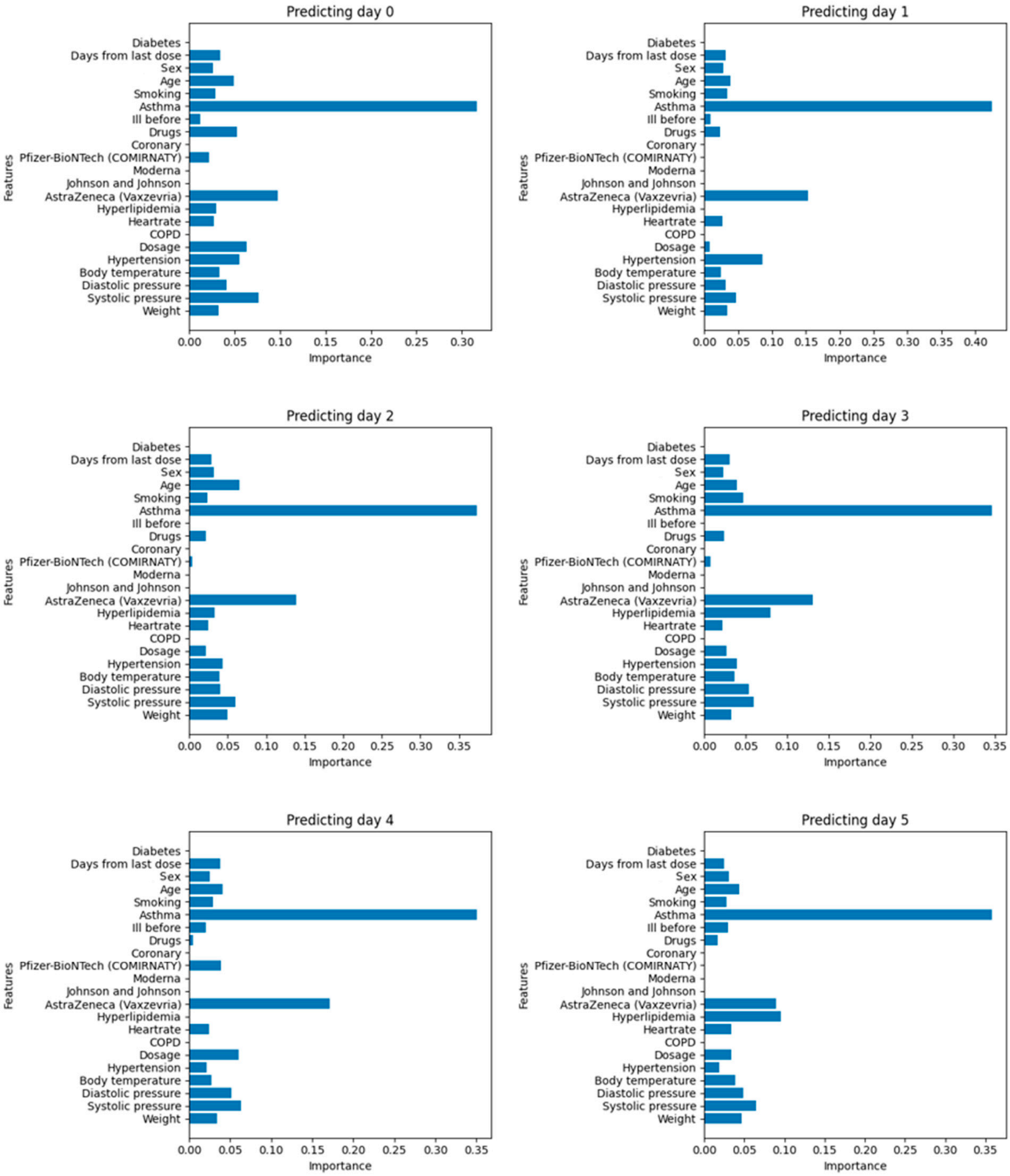


Figure 4. The importance score of variable features of the Extreme Gradient Boosting classifier for predicting the mMRC grade in the coming days.

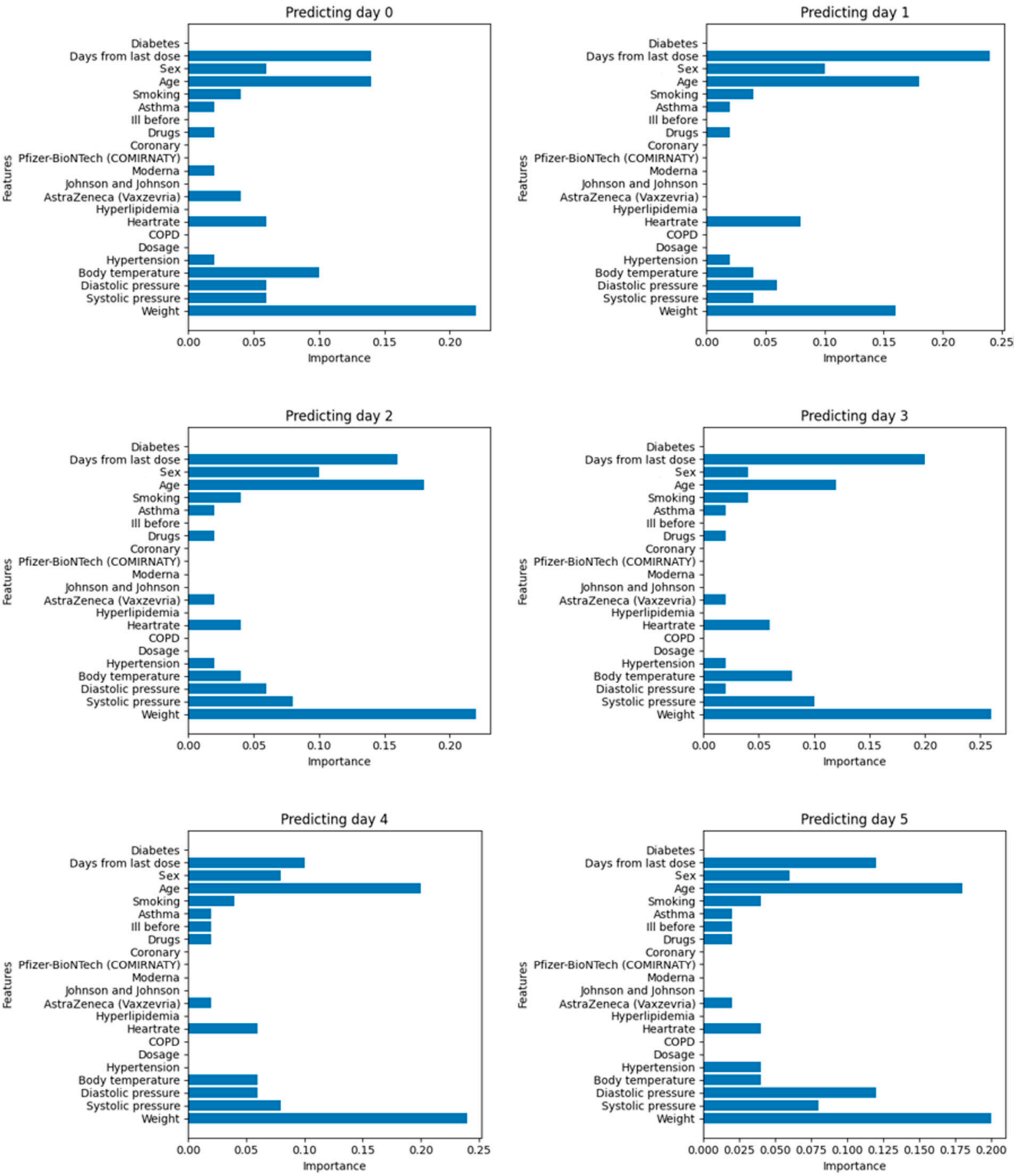


Figure 5. The importance score of feature variables of the AdaBoost classifier for predicting the mMRC grade in the coming days.

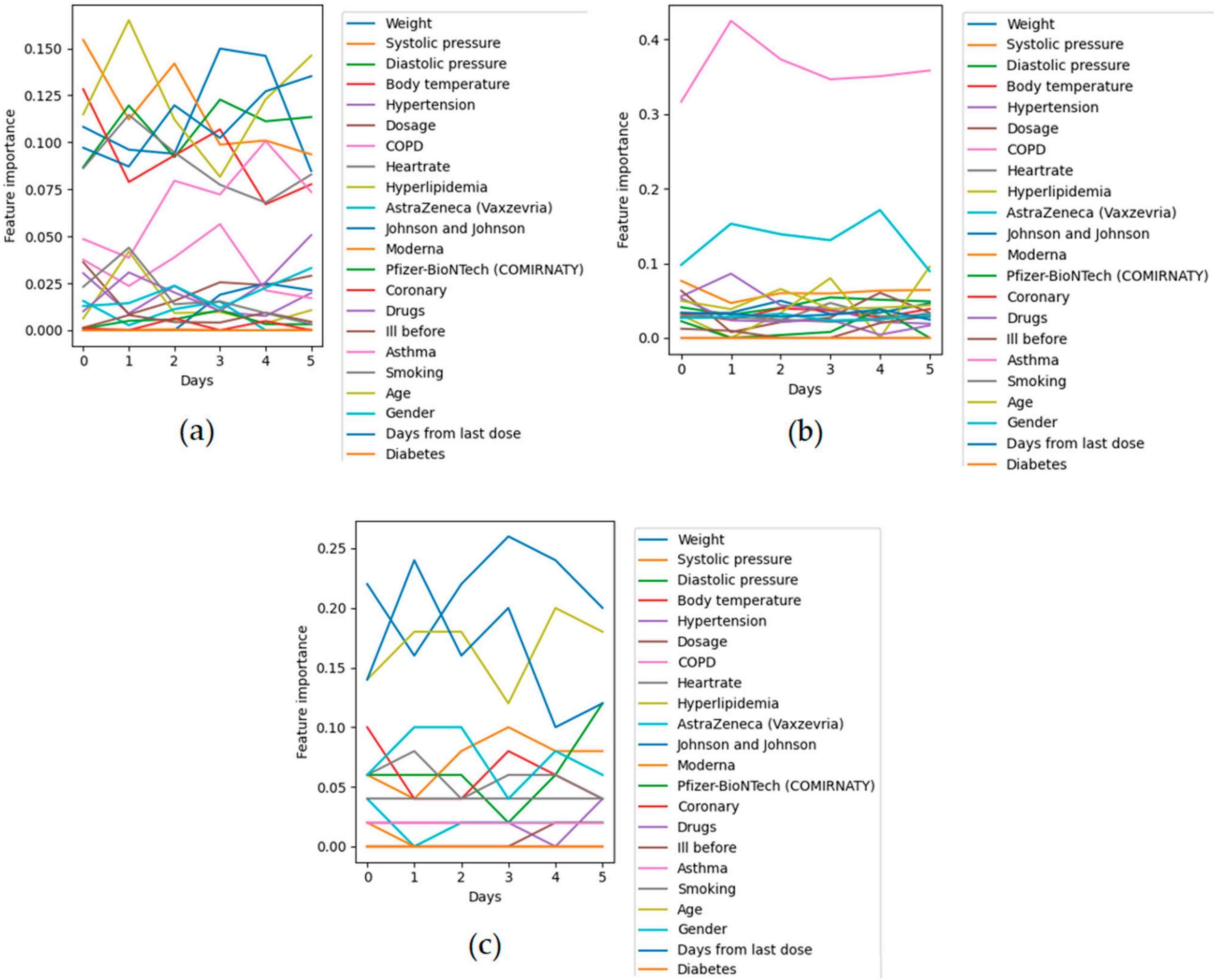


Figure 6. The importance score of variables with respect to days for prediction for the (**a**) Random Forest Classifier, (**b**) XGboost classifier, and (**c**) AdaBoost classifier.

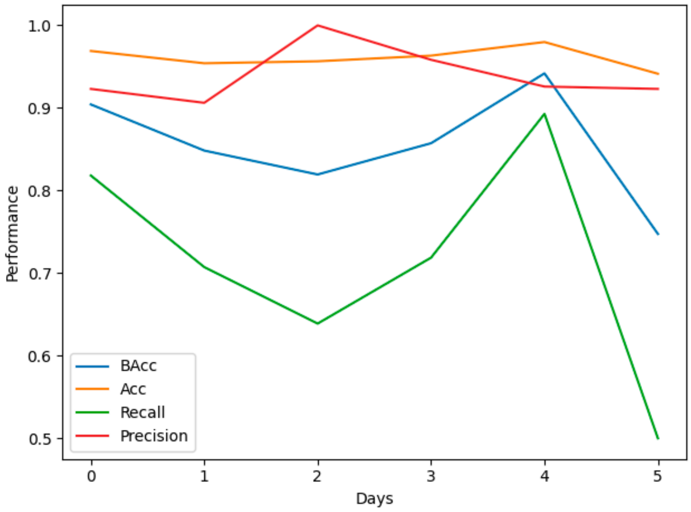


Figure 7. Performance metrics for mMRC classification task by utilizing improved Random Forest classifier.

The classification results are improved compared to those of the original version and show improvement in balanced precision (starting at 0.905) and recall.

3.4. Time-to-Event Analysis for Fever Remission

To predict fever duration, TTE analysis was applied, focusing on the event of fever remission. Patients who did not report fever as part of their initial symptoms on forms did not develop a high body temperature during their follow-up either. For this reason, they are excluded from this analysis, resulting in a new dataset comprising data from 90 patients. To address collinearity [39], which is a common issue when estimating linear or generalized linear models, including Cox regression, and to reduce the dimensionality of the dataset, baseline and daily symptoms that occurred in fewer than five patients (approximately 5% of the dataset) were included in the "Other symptoms" variable. Additionally, in each experiment, daily questionnaires and measurements were aligned with a specified monitoring period, leading to distinct prediction models trained on data filtered to include only patients with fever throughout the specified duration. Consequently, for a follow-up period of *n* days, the analysis was conducted on data of febrile patients obtained from their admission up to *n* days later. The Cox regression model applied for a one-day follow-up analysis was structured as follows: coxph (TTE ~ Demographics + Clinical History + Symptom Variables + COVID-19 Specific Variables + Average Biosignal Measures).

The dependent variable in the analysis, TTE (Time-to-Event), represents the period from their admission to the day of fever remission, capturing the time interval until a patient's body temperature returns to normal levels. The independent variables included demographics and clinical history, which encompassed age, sex, smoking status, asthma, COPD, hypertension, hyperlipidemia, diabetes, coronary artery disease, and medication status. Symptom variables account for initial symptoms and their duration in days, featuring anorexia, cough, headache, fatigue, myalgia, sore throat, and fever, along with days recorded with mMRC grades 1 and 2. COVID-19 specific variables cover the number of vaccine doses received, days since the last vaccine dose, history of prior COVID-19 infection, days since testing positive for COVID-19, and categorization of vaccine manufacturers, such as AstraZeneca, Johnson & Johnson, and Pfizer-BioNTech. Lastly, average biosignal measures include the average values of heart rate, oxygen saturation (SpO_2), body temperature, weight, and diastolic and systolic blood pressure.

From the time-to-event analysis, several noteworthy observations emerged. The days since testing positive for COVID-19 show a pronounced impact, suggesting that as the number of days from testing positive increases, the likelihood of achieving a normothermic state (36.5–37.5 °C) also increases considerably. Similarly, recipients of the Johnson & Johnson vaccine exhibit an increased likelihood of reverting to a normothermic state swiftly. Another notable finding is a marked increase in the duration of high body temperature with every additional day exhibiting the symptom of anorexia. The average oxygen saturation (SpO_2 mean) also shows a potentially impactful trend, though it is just above the common significance threshold. The results from the multivariate Cox regression model applied to the one-day follow-up dataset are presented in Figure 8, which details the ranking of variables based on their log-transformed hazard ratios (log HR).

The model with a one-day follow-up period scores a concordance index of 0.77, while the models with a two-day and three-day follow-up period score a concordance index of 0.82 and 0.94, respectively. The *p*-value for 'Days since positive test' is 0.02 and the hazard ratio (HR) is 17.81 > 1, suggesting a strong relationship between time since infection and increased likelihood of fever resolution. The same conclusion can be drawn for the 'Johnson & Johnson' type of vaccine. In contrast, the *p*-value for 'Symptom days: anorexia' is 0.01 and the proportional hazard (HR) value is 0.32 < 1, suggesting a strong relationship between the number of days with the anorexia symptom and a reduced possibility of fever subsidence. The model with the two-day follow-up period trained on data from a total of 72 patients produced similar results on the characteristic variables days since positive test, Johnson & Johnson, and days with symptom: anorexia. Furthermore, it demonstrates a strong relationship between the variable 'Medication' and a reduced likelihood of fever resolution (p-value = 0.03 and HR = 0.18), as well as a strong relationship between mean systolic pressure and an increased likelihood of fever remission (p-value = 0.03 and HR = 15.32).

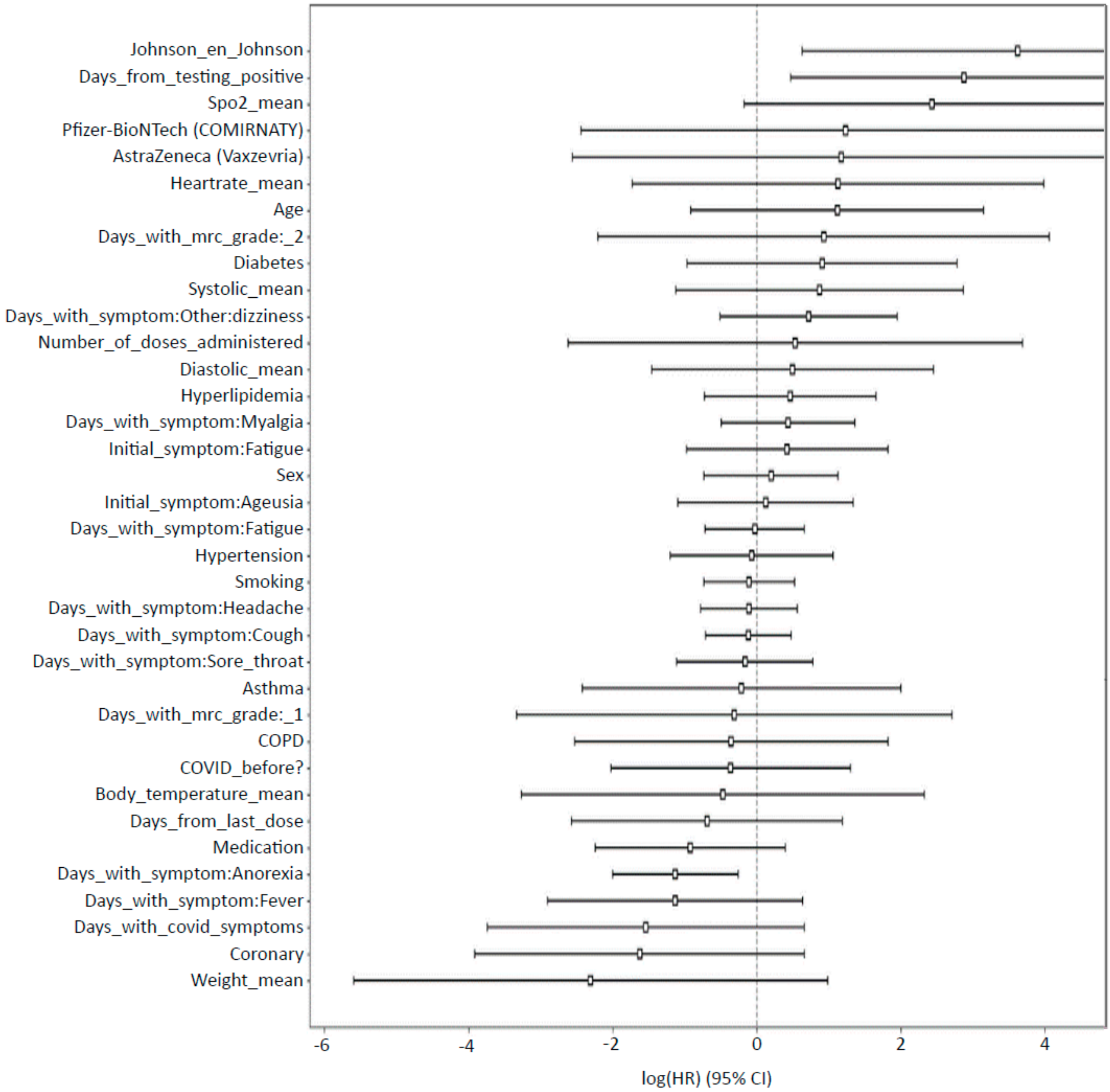


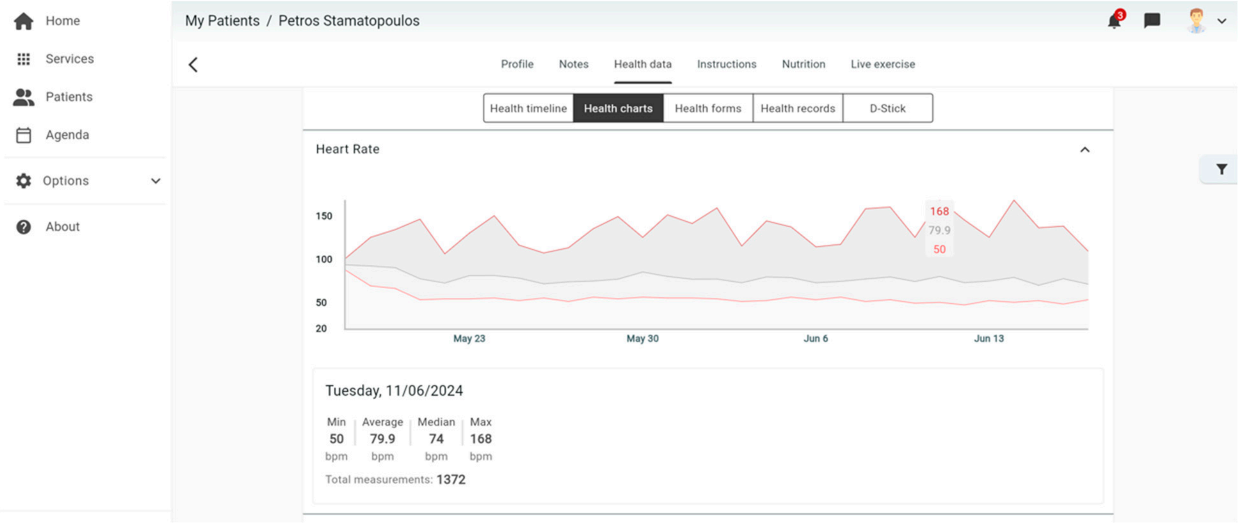
Figure 8. Plot of variable ranking based on log (HR).

The model with a three-day follow-up period, however, returns no features with a p-value $\leq$ 0.05 and includes data from a total of 51 patients. Models with longer follow-up periods than the samples available so far do not converge due to high collinearity, as the number of cases is significantly reduced (fewer than 20 patients).

4. The System in Practice

4.1. Platform for Data Collection

A platform dedicated to the collection of health-related information was utilized in the scope of the project. The platform is an IoT solution integrating vital signs monitoring for COVID-19 patients, with Bluetooth-enabled wearables and a mobile app for daily biosignal measurements and self-assessments. It includes a helpdesk and stores data in a Personal Health Record (PHR) accessible to clinicians. The web app supports remote monitoring using state-of-the-art technology and roles for stakeholders. The provided benefits include care continuity, easy access to healthcare, and increased efficiency for providers. The solution leverages smart devices and non-invasive sensors to monitor physical, physiological, and emotional status, promoting self-management and social engagement. In Figure 9, the web interface of the platform is presented. The health-related data of a patient can be shown by various visualizations providing intuitive summaries for each biosignal.



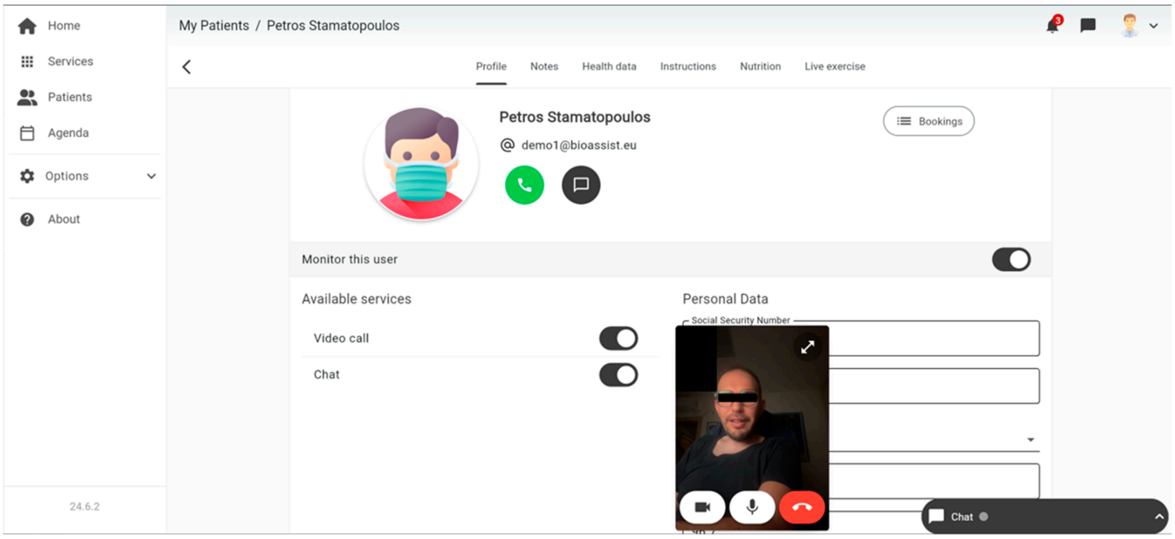
Tuesday, 11/06/2024

Min 50 bpm | Average 79.9 bpm | Median 74 bpm | Max 168 bpm

Total measurements: 1372

Figure 9. Web interface of the platform for data collection.

Through the platform, healthcare experts were able to view patient data (visualization) and conduct virtual visits as depicted in Figure 10.



[redacted]

Petros Stamatopoulos

demo1@bioassist.eu

Monitor this user

Available services

Video call

Chat

Personal Data

Social Security Number

Figure 10. Web interface for conducting virtual visits.

4.2. Platform for Data Analysis

A different web application was developed to integrate the functionality of the proposed methodologies. This architecture was based on the logic of Service-Oriented Architecture (SOA). It consisted of a web interface subsystem and a backend subsystem. The application sends a GET request to the web service, developed in Python through the 'flask' library. The web service can be called at a specific URL. To analyze the mMRC grade classification, the call sent the hyperparameters to the web service in the following format, as an example:

```
GET/daybyday_inference?&cls_type=RF&sex_flag=1.0&days_ahead=3&age_lower_flag=50&age
_upper_flag=80
```

Hyperparameters 'days_ahead' and 'cls_type' were required, while hyperparameters 'sex_flag', 'age_lower_flag', 'age_upper_flag', 'sdatetime_lower_flag', and 'sdatetime_upper_flag' were optional to perform analysis on the filtered samples.

For time-to-event analysis, the call sent the hyperparameters to the web service in the following format, as an example:

```
GET/cox_regression/?follow_up_period_flag=2&sex_flag=1&age_lower_flag=20&initial_positive
_test_date_lower_flag=2020-12-03&initial_positive_test_date_upper_flag=2022-04-30
```

The backend subsystem received the request and transformed the original dataset from the hyperparameters into the corresponding data frames. Figure 11 depicts a Cox regression analysis scenario with an observation period of 2 days on data of male patients over 20 years old who were infected with COVID-19 between the dates of 3 December 2020 and 30 April 2022.

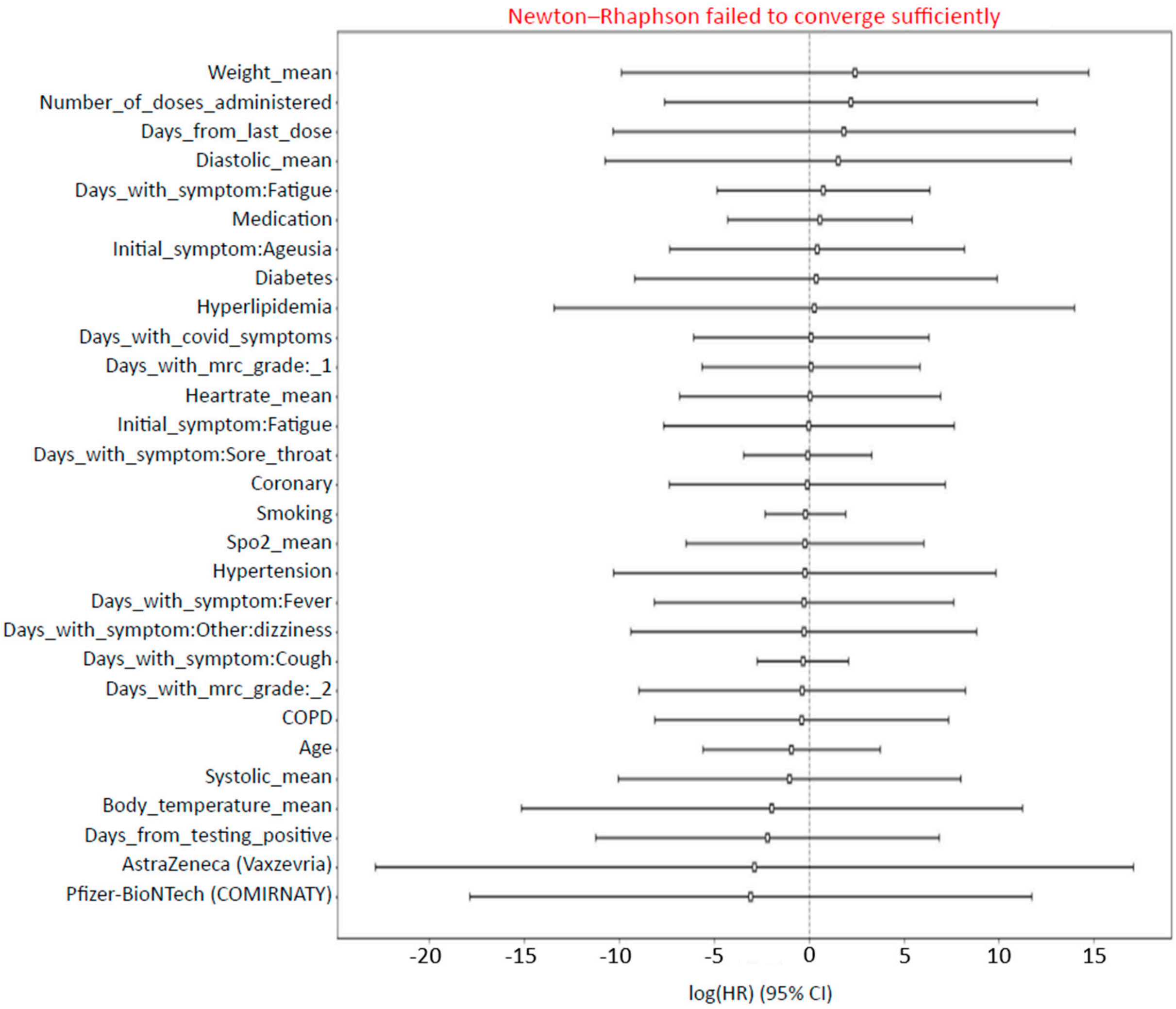


Figure 11. Cox regression analysis with an observation period of 2 days on filtered data of male patients over 20 years old who were infected with COVID-19 between the dates of 3 December 2020 and 30 April 2022.

5. Conclusions

The paper investigates the possibility of extracting useful knowledge from the data of patients with mild COVID-19 symptoms. For this purpose, the use of ML algorithms was examined for the classification of patients on the mMRC scale. The classifiers were selected based on their performance on noisy data and the explanation of their decision-making logic. Therefore, classifiers, which are based on ensemble schemes such as boosting and bagging with base classifiers, and decision trees were considered ideal candidates. This is mainly due to the confirmed results in the relevant literature of increased performance in deviation and dispersion error, and the interpretations of their predictions resulting in transparency and trust in the results produced. Although the dataset showed a significant imbalance in terms of mMRC scale classes, the selection classifiers demonstrated significant robustness as shown by the comparison of metric accuracy and balanced accuracy. At the same time, the overview of the classification results highlights the possibility of using ML classifiers for risk assessment even in groups of patients who maintain small differences in disease progression. However, as can be easily seen, the predictability of the mMRC scale class decreased with time, which is expected. From thorough analysis of the interpretability results, useful conclusions are drawn in relation to the factors influencing the determination of the outcome over time as well as for the filtering of input characteristics. More specifically, classification algorithms agree that factors such as days since the last vaccination, systolic and diastolic blood pressure, weight, age, and heart rate significantly influence the classifiers' decision. In addition, the expected duration of fever is analyzed by applying a Cox regression model. The multivariate Cox proportional hazard model was chosen with the main criteria being the ability to evaluate the parallel effect of several variables as well as the handling of numerical as well as categorical variables. From analysis of the results, useful conclusions emerged in relation to the factors that affect the reduction of fever. The analysis was conducted for three different durations of follow-up periods in patients who manifested an elevated body temperature, and their results showed consistency for two of them. The fact that the results of the analysis with a 3-day follow-up period differ from the results of the other two analyses may be related to the limited number of samples and the problem of overfitting the method to them. Analyses for the one-day and two-day follow-up periods agree that the factors of time of illness (days since positive test), vaccine manufacturer, and time with the anorexia symptom significantly affect the length of time to return to normal body temperature. The results of the classification and regression methods are consistent with the experience of experts, which will work positively in the process of integrating automated risk assessment systems into the daily routine of healthcare delivery structures.

To better inform healthcare professionals and provide personalized healthcare to patients, statistical and ML analysis on tabular data can be successfully applied. Initial data assessment and preparation are a cornerstone for the application of ML techniques that can contribute to personalized information and personalized patient care. More specifically, the use of calibrated scales, such as the mMRC scale, appears to greatly assist in training analysis models and deriving objective results. Additionally, the classification methodology tested showed very good results in terms of classification efficiency against unbalanced datasets and their interpretability properties. Furthermore, the Cox proportional hazards model was applied to estimate the duration of symptoms in patients with infection, producing interesting insights for determining crucial factors for fever remission. Finally, the integration of all of the above into pilot software seems to be feasible and necessary for the automated analysis of health data in order to provide personalized care to patients. Future work focuses on developing and evaluating the effectiveness of a pilot informatics tool that can be integrated into health data analysis systems, providing repeatable, objective, and trustworthy predictions for the progression of mild COVID-19 patients remotely.

Author Contributions: Conceptualization, A.M., C.P. and I.M.; methodology, M.T. and A.K. (Athanasios Kallipolitis); software, M.T. and A.K. (Athanasios Kallipolitis); validation, E.J., P.G., C.P., A.M., I.D., A.K. (Anastasia Kotanidou), I.M. and A.G.V.; formal analysis, M.T., A.K. (Athanasios Kallipolitis), P.G., A.M. and A.G.V.; investigation, E.J., P.G. and A.G.V.; resources, C.P., A.M., I.D., A.K. (Anastasia Kotanidou) and I.M.; data curation, E.J., I.D., A.K. (Anastasia Kotanidou) and A.G.V.; writing—original draft E.J., M.T. and A.K. (Athanasios Kallipolitis); preparation, E.J., M.T., A.K. (Athanasios Kallipolitis) and C.P.; writing—review and editing, P.G., C.P., A.M., I.D., A.K. (Anastasia Kotanidou), I.M. and A.G.V.; visualization, E.J., P.G., C.P., A.M., A.K. (Anastasia Kotanidou), I.M. and A.G.V.; supervision, A.K. (Anastasia Kotanidou) and I.M.; project administration, E.J. P.G., C.P., A.M., A.K. (Anastasia Kotanidou), I.M. and A.G.V.; funding acquisition, C.P., A.M., A.K. (Anastasia Kotanidou) and I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the HORIZON 2020 Programme of the European Commission-H2020-SC1-PHE-CORONAVIRUS-2020-2-CNECT, Grant Agreement number 101016065—COVID eXponential Programme (COVID-X).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the 'Evangelismos Hospital', Athens, Greece (473/7-10-2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are available upon reasonable request.

Conflicts of Interest: Parisis Gallos and Christos Panagopoulos were employed by the company BioAssist SA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Spänig, S.; Emberger-Klein, A.; Sowa, J.; Canbay, A.; Menrad, K.; Heider, D. The Virtual Doctor: An Interactive Artificial Intelligence based on Deep Learning for Non-Invasive Prediction of Diabetes. *Artif. Intell. Med.* **2019**, *100*, 101706. [CrossRef] [PubMed]
2. Gu, D.; Zhao, W.; Xie, Y.; Wang, X.; Su, K.; Zolotarev, O. A Personalized Medical Decision Support System Based on Explainable Machine Learning Algorithms and ECC Features: Data from the Real World. *Diagnostics* **2021**, *11*, 1677. [CrossRef] [PubMed]
3. Qin, F.; Lv, Z.; Wang, D.; Hu, B.; Wu, C. Health status prediction for the elderly based on machine learning. *Arch. Gerontol. Geriatr.* **2020**, *90*, 104121. [CrossRef] [PubMed]
4. Habib, M.; Wang, Z.; Qiu, S.; Zhao, H.; Murthy, A.S. Machine Learning Based Healthcare System for Investigating the Association Between Depression and Quality of Life. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2008–2019. [CrossRef] [PubMed]
5. Latif, S.; Qadir, J.; Farooq, S.; Imran, M.A. How 5G Wireless (and Concomitant Technologies) Will Revolutionize Healthcare? *Future Internet* **2017**, *9*, 93. [CrossRef]
6. Iqbal, S.M.; Mahgoub, I.; Du, E.; Leavitt, M.A.; Asghar, W. Advances in healthcare wearable devices. *NPJ Flex. Electron.* **2021**, *5*, 1–14. [CrossRef]
7. Pradhan, B.K.; Bhattacharyya, S.; Pal, K. IoT-Based Applications in Healthcare Devices. *J. Healthc. Eng.* **2021**, *2021*, 6632599. [CrossRef]
8. Chen, P.; Lin, C.; Wu, W. Big data management in healthcare: Adoption challenges and implications. *Int. J. Inf. Manag.* **2020**, *53*, 102078. [CrossRef]
9. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **2019**, *32*, 18069–18083. [CrossRef]
10. Ranjan, Y.; Rashid, Z.; Stewart, C.L.; Conde, P.; Begale, M.; Verbeeck, D.; Boettcher, S.; Dobson, R.J.; Folarin, A.A. RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR mHealth uHealth* **2019**, *7*, e11734. [CrossRef]
11. Bhat, G.; Deb, R.; Ogras, U.Y. OpenHealth: Open-Source Platform for Wearable Health Monitoring. *IEEE Des. Test* **2019**, *36*, 27–34. [CrossRef]
12. Bahmani, A.; Alavi, A.; Buergel, T.; Upadhyayula, S.; Wang, Q.; Ananthakrishnan, S.K.; Alavi, A.H.; Celis, D.; Gillespie, D.; Young, G.; et al. A scalable, secure, and interoperable platform for deep data-driven health management. *Nat. Commun.* **2021**, *12*, 5757. [CrossRef] [PubMed]
13. Hermes, S.; Riasanow, T.; Clemons, E.K.; Böhm, M.; Krcmar, H. The digital transformation of the healthcare industry: Exploring the rise of emerging platform ecosystems and their influence on the role of patients. *Bus. Res.* **2020**, *13*, 1033–1069. [CrossRef]
14. Najim, A.H.; Elkhediri, S.; Alrashidi, M.; Nasri, N. The Impact of using IoT for Elderly and Disabled Peoples Healthcare: An Overview. In Proceedings of the 2022 2nd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 25–27 January 2022; pp. 394–398.
15. Menychtas, A.; Tsanakas, P.; Maglogiannis, I. Knowledge discovery on IoT-enabled mHealth applications. In *GeNeDis 2018: Computational Biology and Bioinformatics*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 181–191.

16. Kyriazis, D.; Autexier, S.; Boniface, M.; Engen, V.; Jimenez-Peris, R.; Jordan, B.; Jurak, G.; Kiourtis, A.; Kosmidis, T.; Lustrek, M.; et al. The CrowdHEALTH project and the hollistic health records: Collective wisdom driving public health policies. *Acta Inform. Medica* **2019**, *27*, 369. [CrossRef]
17. Dinh-Le, C.; Chuang, R.; Chokshi, S.K.; Mann, D.M. Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions. *JMIR mHealth uHealth* **2019**, *7*, e12861. [CrossRef]
18. Rawajbeh, M.A.; Band, S.S.; Mosavi, A.H.; Kumar, R.L.; Khan, F.; Din, S.; Ibeke, E. Recurrent Neural Network and Reinforcement Learning Model for COVID-19 Prediction. *Front. Public Health* **2021**, *9*, 744100.
19. Kallipolitis, A.; Gallos, P.; Menychtas, A.; Tsanakas, P.; Maglogiannis, I. Medical Knowledge Extraction from Graph-Based Modeling of Electronic Health Records. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, León, Spain, 14–17 June 2023; Springer: Cham, Switzerland, 2023.
20. Yuan, Z.; Ding, H.; Chao, G.; Song, M.; Wang, L.; Ding, W.; Chu, D. A Diabetes Prediction System Based on Incomplete Fused Data Sources. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 384–399. [CrossRef]
21. Cutillo, C.M.; Sharma, K.R.; Foschini, L.; Kundu, S.; Mackintosh, M.; Mandl, K.D.; MI in Healthcare Workshop Working Group. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digital Medicine* **2020**, *3*, 47. [CrossRef]
22. Rueda, J.; Rodríguez, J.D.; Jounou, I.P.; Hortal-Carmona, J.; Ausín, T.; Rodríguez-Arias, D. "Just" accuracy? Procedural fairness demands Intepretability in AI-based medical resource allocations. *Ai Soc.* **2022**, 1–12.
23. Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2019**, *48*, 137–141. [CrossRef]
24. Tziomaka, M.; Kallipolitis, A.; Tsanakas, P.; Maglogiannis, I. Evaluating Mental Patients Utilizing Video Analysis of Facial Expressions. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Crete, Greece, 25–27 June 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 182–193.
25. Reza Soroushmehr, S.M.; Najarian, K. Transforming big data into computational models for personalized medicine and health care. *Dialogues Clin. Neurosci.* **2016**, *18*, 339–343. [CrossRef]
26. Fröhlich, H.; Balling, R.; Beerenwinkel, N.; Kohlbacher, O.; Kumar, S.; Lengauer, T.; Maathuis, M.H.; Moreau, Y.; Murphy, S.A.; Przytycka, T.M.; et al. From hype to reality: Data science enabling personalized medicine. *BMC Med.* **2018**, *16*, 150. [CrossRef]
27. Lonergan, M.; Senn, S.J.; McNamee, C.; Daly, A.K.; Sutton, R.; Hattersley, A.T.; Pearson, E.R.; Pirmohamed, M. Defining drug response for stratified medicine. *Drug Discov. Today* **2017**, *22*, 173–179. [CrossRef]
28. Evans, S.S.; Repasky, E.A.; Fisher, D.T. Fever and the thermal regulation of immunity: The immune system feels the heat. *Nat. Rev. Immunol.* **2015**, *15*, 335–349. [CrossRef]
29. Launey, Y.; Nesseler, N.; Mallédant, Y.; Seguin, P. Clinical review: Fever in septic ICU patients–friend or foe? *Crit Care* **2011**, *15*, 222. [CrossRef]
30. Huang, T.; Guo, Y. Application and effects of fever screening system in the prevention of nosocomial infection in the only designated hospital of coronavirus disease 2019 (COVID-19) in Shenzhen, China. *Infect. Control Hosp. Epidemiol.* **2020**, *41*, 978–981. [CrossRef]
31. Mahler, D.A.; Wells, C.K. Evaluation of clinical methods for rating dyspnea. *Chest* **1988**, *93*, 580–586. [CrossRef]
32. Silverman, B.W.; Jones, M.C.E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *Int. Stat. Rev./Rev. Int. De Stat.* **1951**, *57*, 233–238. [CrossRef]
33. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
34. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
35. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. [CrossRef]
36. Alshayeji, M.H. Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1195–1213. [CrossRef]
37. Cox, D.R. Regression Models and Lifetables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–202. [CrossRef]
38. Emmert-Streib, F.; Dehmer, M. Introduction to Survival Analysis in Practice. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 1013–1038. [CrossRef]
39. Xue, X.; Kim, M.Y.; Shore, R.E. Cox regression analysis in presence of collinearity: An application to assessment of health risks associated with occupational radiation exposure. *Lifetime Data Anal* **2007**, *13*, 333–350. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.