Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 1232–1241

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Machine Learning Models for Predicting Short-Long Length of Stay of COVID-19 Patients

Matteo Olivato[a], Nicholas Rossetti[a], Alfonso E. Gerevini[a], Mattia Chiari[a], Luca Putelli[a], Ivan Serina[a]

[a] *Università degli Studi di Brescia, Via Branze 38, Brescia, Italy*

Abstract

During 2020 and 2021, managing limited healthcare resources and hospital beds has been a fundamental aspect of the fight against the COVID-19 pandemic. Predicting in advance the length of stay, and in particular identifying whether a patient is going to stay in the hospital longer or less than a week, can provide important support in handling resources allocation. However, there have been significant changes in terms of containment measures, virus diffusion, new treatments, vaccines, and new variants of SARS-CoV-2 during the last period. These changes pose several conceptual drift issues that can limit the usefulness of machine learning in this context. In this work, we present a machine learning system trained and tested using data from more than 6000 hospitalised patients in northern Italy, distributed over almost two years of pandemic. We show how machine learning can be effective even by analysing data over this long period of time, also exploiting a model that predicts the patient's outcome in terms of discharge or death. Furthermore, learning from data that also consider deceased patients is a common issue in predicting the length of stay because they have severe conditions similar to patients with a long stay period, but may actually have a very short duration of hospitalisation. For this purpose, we present a method for handling data from alive and deceased patients, exploiting more patient records, increasing the robustness of the model and its performance in this task. Finally, we investigate the features that are most relevant to the prediction of the simplified length of stay.

1. Introduction

For more than two years, the scientific community has made great efforts to counter the spread and harmful effects of COVID-19, which reached more than 300 million cases in January 2022. AI and ML communities are developing models and tools to tackle the epidemic at various levels [7, 8]. These include diagnosis, prognosis [11, 13] and the discovery of treatments and tools to help health facilities during an epidemic, such as managing limited healthcare resources to avoid hospital overloads and supporting physicians' decisions.

[*] Matteo Olivato; m.olivato@unibs.it

10.1016/j.procs.2022.09.179

This last aspect is the focus of our work, which analyses machine learning techniques to assess a simplified length of stay for hospitalised patients with COVID-19. Patient data are collected in *Spedali Civili di Brescia*, one of the hospitals that had more COVID-19 patients in Italy. These data include demographic information, such as sex and age, ten different laboratory test values related mainly to blood values recorded at different times during hospitalisation, an indication of the eventual hospitalisation in an intensive care unit, and the number of the hospital ward in which the patient is hospitalised.

Although many studies focus on data belonging to a limited period of time [1, 4, 30], we have access to more than 6000 patients from March 2020, on the most critical days of emergency, until the end of 2021. During this time period, important changes, such as the prevalence of Alpha and Delta SARS-CoV-2 variants, the different diffusion of the virus (and its effect on hospitalisations), the introduction of vaccines and new treatments, can lead to significant concepts drift issues and limit the usefulness of machine learning techniques.

Providing an initial assessment of the patients' conditions can be very useful in better managing medical resources, especially during the most severe periods of the epidemic. Therefore, we address the task of predicting whether a patient will stay longer or less than a week in the triage phase, taking into account several machine learning techniques (Ensemble of Decision Trees, Boosting algorithms, and Feed-Forward Neural Networks). In addition, we show how the integration of the outcome (in terms of release or death) predicted by another machine learning model can also improve the evaluation of the length of stay.

Real-world data can present several quality issues, such as a limited number of features, missing values, and noise, especially in the clinical domain [14, 24] and during emergency conditions [12, 16]. In our context, medical treatments and comorbidities were not available among the collected data. Despite these limitations, we show that machine learning techniques can also be effective in this challenging domain. In fact, our models reported good performance on a binary classification task based on length of stay prediction, identifying patients who will stay longer than a week.

In the following, after briefly presenting the machine learning algorithms adopted, we consider and discuss related work, and present our architecture. Then we report our experimental evaluation, and finally we give our conclusions and mention future work.

2. Background and Related work

2.1. Machine learning algorithms

In this section, we briefly introduce the machine learning algorithms used in our Simplified Length of Stay (SLOS) prediction system.

Random Forests (RF) [5] is an ensemble learning method that builds a number of decision trees at training time. To build each individual tree of the random forest, randomly chosen subsets (with replacement) of the data features and of the training samples are used. In our implementation, the output value is obtained by averaging the values provided by all trees. On the other hand, Extremely Randomised Trees (Extra Trees or ET) [15] are another ensemble learning method based on decision trees. The main differences between Extra Trees and Random Forests are the following. In the standard Decision Tree used by Random Forest, the cut point is chosen by first computing the optimal cut point for each feature, and then choosing the best feature to branch the tree; in Extra Trees, for each tree, the algorithm randomly chooses k features and then, for each chosen feature f, it randomly selects a cut point C_f in the range of possible f values. This generates a set of k couples $\{(f_i, C_i) \mid i = 1, \dots, k\}$. Then, the algorithm compares the splits generated by each couple (e.g., under the split test $f_i \le C_i$) to select the best split using a quality measure, such as the Gini index.

The Gradient Boosting Method (GBM) incrementally assembles several *weak learners*, such as Decision Trees, to produce a single predictive model. GBM uses a gradient descent procedure to assign a weight to each weak learner, whose value is related to the learner's ability to reduce errors. XGBoost (eXtreme Gradient Boosting) [9] is one of the most popular GBM algorithms that uses Decision Trees as a weak learner. An important aspect of this algorithm is how it controls overfitting, which is a known issue in gradient boosting algorithms. XGBoost adopts a more regularised model formalisation, which allows it to obtain better results than GBM. Another popular GBM algorithm is LightGBM [18]. It adopts certain differences with respect to XGboost, such as, for instance, replacing continuous values with discrete bins, building each weak learner according to samples that actually have more impact on the loss function, and growing the internal trees leaf-wise (node-by-node).

A Feed-Forward Neural Network consists of layers of artificial neurons that form a directed acyclic graph with weighted edges. These models are made up of an input layer, one or more hidden layers, and an output layer. The input features that make up the input layer are connected to a first hidden layer composed of a series of neurons. Each neuron computes the weighted sum of all inputs using the weights of the input edges, and then a bias term is added. An activation function is applied to the result that produces the output of the neuron. The output of a hidden layer, which is the array of the outputs of all neurons in that layer, can be connected to another hidden layer or to the output layer. The output layer provides the predicted class or the regression value for the input training instances. The learning algorithm compares the predictions with the target values and evaluates the loss function of the model. During the training phase, a Feed-Forward Neural Network learns the weights of each layer by minimising the loss function through an optimisation algorithm that is typically Back-Propagation.

2.2. Related work

Artificial Intelligence and Machine Learning have been used in the medical domain in contexts such as radiology [14, 24], analysis of clinical documents [23], and pharmacology research [25]. Several surveys, such as the works in [3, 26], present an overview of recent studies on the use of Artificial Intelligence against COVID-19 for drug discovery and development, testing and diagnosis (especially via X-Ray imaging), tracking and epidemiology, and prediction of patient outcome. Furthermore, an overview of the issues and challenges of applying ML in a critical care context is available in [16]. This work stresses the need to deal with corrupted data, such as missing values, imprecision, and errors, which can increase the complexity of prediction tasks.

Despite these issues, Abdulaal et al. [1] propose a Neural Network model for mortality prediction that is trained and tested using data from approximately 400 patients in the United Kingdom. Unlike our work, this study does not consider any laboratory tests performed at the beginning of hospitalisation; instead it uses information obtained when a patient is admitted to a hospital (such as symptoms, demographic information, and smoking history). The work in [4] investigates the tasks of predicting mortality within 14 or 30 days after diagnosis. They trained their models with 10000 patients in Korea, which is almost twice our number of patients, but only patients from January to April 2020, focused on the initial emergency period of the pandemic. In their article, the authors report results using LASSO, Support Vector Machines and Random Forest, and they obtained the best performance using LASSO and Linear SVM. On the other hand, the problem of predicting the length of stay of patients from a statistical point of view is the focus of [29] and [19]. The first focusses on English patient data collected in the National Surveillance System (CHESS) database and uses techniques such as the Accelerated Failure Time and Truncation Correction Method. They only identified a few features as predictors, such as sex, age, and week of admission, without considering any laboratory test or vaccination information. The second work describes a probabilistic tool made in Singapore that combines the prediction of the length of stay with the fatality rate, hospital capacity, and other features to build a bed resource plan in the best-case, base-case and worst-case scenarios. It uses only data from April 2020 to project the need for bed resources in the next month, May 2020.

Machine learning techniques are used in [22] to predict length of stay, together with a survival analysis, for a dataset made up of more than 1000 patients. They consider features related to symptoms, infection and hospitalisation dates, travel and chronic diseases histories, death or discharge status, and even demographics. On the contrary, they do not consider laboratory tests or X-ray scores. The authors compare the results of traditional statistic methods with Support Vector Machines and Gradient Boosting algorithms, dealing with missing values and no comorbidities due to emergency conditions.

We also published some related work on COVID-19 tasks. Decision tree ensembles are the main model used by Gerevini et al. [13] to predict the outcome (alive or decease) at different times during the stay on data from 2000 patients in Northern Italy. This study monitors the progression of COVID-19 considering laboratory tests and their trends in terms of improvement, stable conditions or worsening and introduces an innovative technique to recognise less reliable predictions made by the system. Furthermore, we introduced an outcome prediction approach based on Recurrent Neural Networks in [11] on an extended cohort of patients. Using GRU units, we achieved interesting performance in patient hospitalisations data considered as time series. Furthermore, this particular representation allows us to use a data augmentation approach to better address the scarcity of samples in both classes.

In terms of predicting the length of stay, our previous work in [10] proposes an analysis of ensembles of decision tree models to predict the length of stay of 1000 COVID-19 patients at different times during their hospitalisation. We

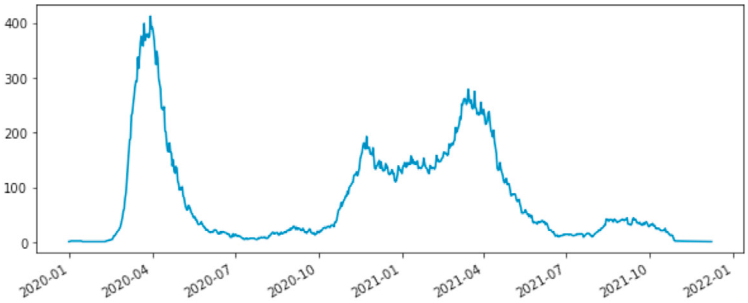


Fig. 1: Hospitalised patients with COVID-19 from January 2020 to December 2021

created and evaluated several datasets considering basically the same characteristics as used in our previous works. There were many issues in the data considered, such as the low number of patients, the lack of comorbidities, and the lack of information on ICU and vaccination, which significantly increased the difficulty of the task. Furthermore, we did not analyse more advanced methods (mainly due to the lack of data), such as boosting or neural networks, focussing only on hospitalisation periods from the initial pandemic.

3. Available Data Sources and Dataset Creation

In this work, we consider data from a total of more than 6, 000 hospitalised patients from March 2020 to December 2021, provided by working in collaboration with the *Spedali Civili di Brescia* Hospital. In Figure 1 we report the trend of daily admission of patients with COVID-19 to our hospital. Between March and May 2020, the first wave of the epidemic is clearly recognisable. While during the summer there was a significant decrease in virus circulation, a second wave hit Italy and our region in the fall (Alpha variant). A third wave, mostly dominated by the Delta variant, occurred between February and April 2021. Despite the fact that the situation has improved drastically with the progress of the vaccination campaign started by the Italian government in the first months of 2021, we can see a small number of hospitalisations (mainly caused by the Delta variant) between September and October.

During hospitalisation, medical personnel performed several exams to monitor patients' conditions, check the response to some treatments, verify the need to transfer a patient to the ICU, etc. For each of these patients, in addition to age and sex, the lab test features that were made available to us are the following:

- the values and dates of several laboratory tests: PCR, LDH, Ferritin, Troponin-T, WBC, D-dimer, Fibrinogen, Lymphocytes and Neutrophils;
- the values and dates of the throat swab exams for COVID-19;
- the identifier of the hospital ward in which the patient is admitted;
- a boolean value (a flag) whether the patient has been admitted to the Intensive Care Unit (ICU) or not;
- the patient's vaccination status, represented as the identifier of the vaccine and the number of received doses;
- the final outcome of hospitalisation at the end of the stay (either in-hospital death, released survivor or transferred to another hospital or rehabilitation centre);
- the length of stay (LOS), calculated as the number of days between the date of admission and the date of release.

We did not have additional information on symptoms, their timing, comorbidities, generic health conditions, virus variants, or clinical treatment. Furthermore, the available data on whether a patient was (or had been) in the ICU and in the admission department present some issues. In fact, during the first wave, practically the entire hospital was involved in the emergency response. Therefore, while a patient is usually treated in a ward related to his/her pathology, during this period there was no formal distinction between wards and their names were only a geographical indication. In terms of the ICU, the limited number of units available was one of the most critical aspects in the early phases of the epidemic, with many patients who could not receive adequate treatment or as long as they needed. As a result, the data that record admission to the ICU in the first wave and in the most severe period of the second wave can be incomplete, noisy, and incoherent.

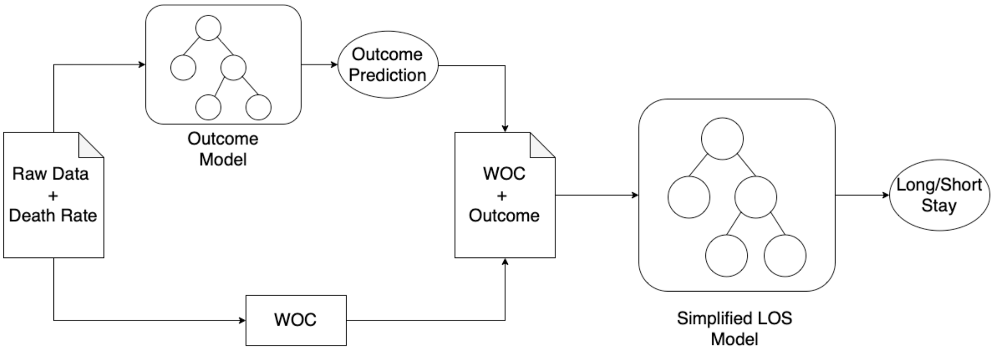


Fig. 2: Architecture of our system. First of all, additional features (the death rate and the estimated probability of the predicted outcome obtained by the Outcome Model) are added. Then the WOC approach is applied to the labels of the deceased patients. On these data, we train the model to predict if the patient is going to stay more than or less than a week.

As in other biomedical applications with raw real-world data [12], other non-trivial practical issues related to the quality of available data are also present for laboratory tests, especially for those patients who were hospitalised in the period of the highest emergency. In our case, the most relevant issue is related to the irregular frequency of tests and exams during the patient's hospitalisation. This is the result of different types and timing of the relative procedures, the need for different resources (X-ray machines, lab equipment, technical staff, etc.), or the different severity of the health conditions of the patients. For these reasons and especially in the triage or even on the first days of hospitalisation, if a test has not been performed, we could have several *missing values*.

As described above, in 2021, the Italian government launched a massive vaccination campaign against COVID-19[1]. Given the effectiveness of vaccines against the risk of hospitalisation, only a small number of vaccinated patients have been treated in *Spedali Civili di Brescia* and according to hospital staff for health issues that were very rarely related to COVID-19. During an epidemic and over a long period of time, such as in our datasets, with many differences in terms of virus circulation, therapies, and clinical knowledge, there can be important changes in the prognostic outcome of hospitalised patients. In machine learning, this phenomenon is known as *concept drift* [27, 33] and can have a negative impact on the performance of learning algorithms. As we explained in [11], since subsampling methods, selecting the most suitable training set according to data distribution, have the disadvantage of reducing training data and potentially leading to poor predictive results, we decided to follow a different approach, using the entire dataset with an additional feature that helps the learning algorithm discriminate whether or not a patient was hospitalised during a particularly critical phase. The new feature, called **death rate**, aims to provide an indicator of the state of pandemic emergency on a given day (when the feature is evaluated) and is defined as the average death rate calculated considering the seven days preceding such a day. Specifically, the death rate feature is the ratio of all patients who died over all patients discharged (dead or alive) during the 7-day period considered.

Our datasets are built considering the first lab tests performed at triage time and also including demographics, the hospital ward, if the patient has been treated in ICU and the vaccination status. We used stratified sampling to select 80% *of the patients for training* the models and 20% *for testing* them.

4. Simplified Length of Stay Prediction Models

Although typically the prediction of length of stay is considered a regression task, the multiplicity of internal (lab test findings, demographic information) and external (available ICU units, hospital overload, etc.) factors that influence the length of stay of a patient during an epidemic makes this regression task very challenging, especially at triage time. However, at this stage, following hospital requests, it could be very useful to provide an approximate indication of the severity of the patient's conditions.

[1] Detailed information can be found at https://github.com/italia/covid19-opendata-vaccini

For this reason, the task we are going to analyse is a simplified version of the Length Of Stay (LOS) prediction, consisting of the binary classification of the duration of hospitalisation (Simplified Length of Stay or **SLOS**). More specifically, we predict whether a patient will have:

- **Short Stay**, if the patient stays in the hospital for 7 days at most;
- **Long Stay**, if the patient stays in the hospital for more than 7 days.

This first simplified approach (the base SLOS model) consists of training a machine learning model to predict whether a patient will have a long or short stay, using the features described in Section 3 for all patients.

A drawback of this approach is that considering the length of stay regardless of the patient's outcome can be confusing. In fact, a patient who dies after just a few days could have the same LOS as a patient with a mild condition who is released after some minor treatments. Therefore, following [10, 6] we also trained and tested the same algorithms considering only patients who are going to be released alive (we call it the Alive Only model) as a second approach. However, the application of this model presents some problems. In fact, an algorithm trained only on patients who are going to be released alive could perform poorly on patients who have a serious death risk. As a result, we propose the Worst Outcome Censoring (WOC) method, which adapts the approach in [6] to our SLOS task. This method uses not only patients released alive, but also those who died during their stay. This technique, compared to the basic approach, evaluates the patient's death as the worst outcome and assigns the maximum possible value to this event; therefore, in our task, we will assign the value of Long Stay to all the dead patients. The intuition behind this approach is that if the patient survives after a worsening of his/her condition, he/she will have to recover from the disease and, therefore, will have a long hospital stay.

As shown in [11, 13] the estimation of the patient's outcome can be very important for the management of health care resources. Following the approach that we adopted and described in [13], we trained another model for this purpose. This **Outcome** model performs a classification task with simply two classes: *Alive* and *Dead*. Although this information can be very useful per-se, it could also be combined with our models to improve their performance.

In Figure 2 we show the architecture of our complete model that takes advantage of the Worst Outcome Censoring and the Outcome Prediction. The main procedure can be described in summary as follows:

- for each patient, we consider his laboratory findings, demographics (age and sex), hospital ward, ICU flag, vaccination status, and considering the patient's admission date, we add the death rate calculated on the previous seven days;
- we train the Outcome Model. This model outputs the estimated probability of the *Alive* and the *Dead* class;
- we enrich our data with the estimated probability of the *Dead* class as an additional feature;
- these enriched data are used to train the SLOS model. When using WOC techniques, the labels of all dead patients are set to *Long Stay*.

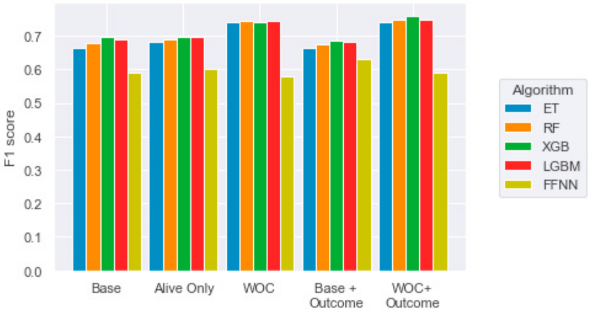
The Outcome and the SLOS models (in all the different versions) are made up of two main components. The first is *Iterative Imputer with Bayesian Ridge Regression* [28] to handle missing values.

With this technique, a Bayesian Ridge Regression algorithm is trained on this new dataset and used to predict the missing values of the output feature. At each step, all features, including those imputed in the previous steps, are used to train the regressor. The second component is the actual machine learning algorithm. As learning algorithms, we considered ensembles of Decision Trees with bagging (Random Forests, ExtraTrees), boosting (XGBoost, LightGBM) techniques and Feed-Forward Neural Networks.

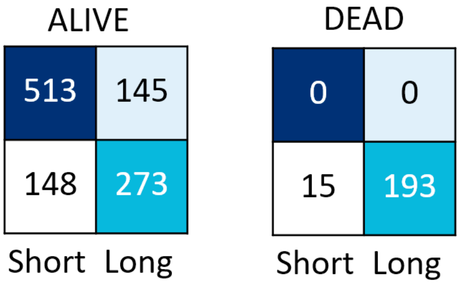
4.1. Training and Hyperparameter Tuning

The training phase of our algorithms is optimised on the *F-β* score metric, which is the weighted harmonic mean of the *precision* and *recall* measures. Specifically, it is defined as follows:

$$F\text{-}\beta = (1 + \beta^2) \cdot \frac{precision + recall}{\beta^2 \cdot precision + recall}$$



(a) Comparison of the results obtained by the different approaches on the same test set in terms of F1 score.


ALIVE

513	145
148	273
Short	Long

DEAD

0	0
15	193
Short	Long

(b) Confusion matrices for patients who were released alive (left) and for the patients who died during the hospitalisation (right) obtained by our best model (WOC + Outcome) on the test set.

However, there are some differences between the SLOS and the Outcome models. Although SLOS is optimised with the common version of the F score (with $\beta = 1$), the Outcome model considers $\beta = 2$. Given that this parameter indicates how many times recall is more important with respect to precision, we chose $\beta = 2$ to minimise false negatives, which are those patients whose very severe conditions are not identified by the algorithm. In fact, these are the most undesirable mistakes, and a decision support system should be optimised to avoid them as much as possible, especially at the triage stage. While the common implementations of the bagging algorithms based on Decision Trees allow one to maximise the F-β score metric, with boosting algorithms and Feed-Forward Neural Networks we require to minimise a loss function. Therefore, following the approach described in [17] we designed a loss function based on this metric, as described by the loss formula: $\mathbb{L} = 1 - F$-β.

It is well-known that the performance of most machine learning algorithms strongly depends on the settings of their hyperparameters. Therefore, to obtain the best performance, the hyperparameters are tuned (automatically) through experiments that evaluate different hyperparameter configurations. We used two different hyperparameter tuning methods: for ensembles of decision trees, we used Random Search; on the other hand, for Neural Networks we used the Bayesian optimisation approach via the Optuna framework [2] with 1024 search iterations. In fact, the need for a faster training phase led us to choose a hyperparameter search method for neural networks that was different from Random Search. Decision tree ensembles can better exploit parallel executions in our systems. Therefore, the Random Search approach, in combination with fast training algorithms, allows us to maximise parallel execution, obtaining good hyperparameter settings in a limited time, despite the high number of considered hyperparameter configurations (4096). In contrast, an optimisation approach for hyperparameters tuning that adapts to the already evaluated configurations puts some limitations in terms of a highly parallel execution. On the other hand, the training time for a single configuration of hyperparameters of a neural network and its initialisation overheads (weights initialisation, data transfer) is significantly higher than for models based on decision trees. Therefore, a smarter approach, such as Bayesian optimisation, allows us to obtain better hyperparameter settings with fewer trials, requiring nearly the same amount of time as the Random Search. For both approaches, we perform hyperparameter tuning using a k-fold cross-validation setup, with $k = 10$, and using the F1 score as the optimisation metric. This validation setup increases the robustness of the best hyperparameter combination found, reducing overfitting of the architecture.

5. Experimental Analysis

As reported in Section 4, we perform a binary classification task to predict whether a patient will stay in the hospital for at most a week (Short Stay) or more (Long Stay). In addition to this task, we use a model to predict the patient's outcome to provide an additional feature to our models. The best Outcome Model for each task is made by an optimised LGBM model based on the F2 score, reaching 70% on that metric and 81% in terms of ROC-AUC.

In this section, we report the experimental results of our models. In summary, the analysed approaches are as follows:

- Base: the models are trained considering all the original patient's data (both alive and decease);

Table 1: Results obtained by our models on the SLOS task in terms of F1, F2, ROC-AUC and Accuracy. In the Short and Long columns, we report the number of patients who stay in the hospital for less than and more than a week, respectively.

Method	Short	Long	Accuracy	Precision	Recall	F1	F2	ROC-AUC
Base	771	516	0.70	0.70	0.70	0.69	0.65	0.70
Alive Only	658	421	0.72	0.71	0.72	0.71	0.72	0.72
WOC	658	629	0.75	0.75	0.75	0.75	0.73	0.75
Base + Outcome	771	516	0.69	0.69	0.69	0.68	0.69	0.69
WOC + Outcome	658	629	**0.76**	**0.76**	**0.76**	**0.76**	**0.74**	**0.76**

- Alive Only: contains the same data as Base, but the models consider only patients released alive;
- WOC: contains the same data as Base, but the labels are changed according to the Worst Outcome Censoring technique;
- Base + Outcome: contains the same data as Base, plus the prediction of the outcome calculated by the Outcome Model as an additional feature;
- WOC + Outcome: contains the same data as Base + Outcome but the labels are changed with the WOC method.

For all these configurations, we considered several machine learning algorithms, i.e. LightGBM, XGBoost, Random Forest, Extra Trees and Feed Forward Neural Networks, increasing their performance through hyperparameter tuning. Then we evaluated them on an isolated test set (composed of 20% of the patients) in terms of the F1 score.

In Figure 3a we can see a more detailed comparison between the performance of the algorithms in different configurations. We can notice how the models based on the ensemble of decision trees outperform the Feed-Forward Neural Networks (FFNN) in all cases. In particular, the boosting algorithms (XGB and LGBM) achieve slightly higher scores in the Alive Only and WOC approaches than the other ones (RF and ET).

The detailed results of the different models are shown in Table 1. We can see how the Base model obtains quite good results in terms of the F1 score and ROC-AUC with values of 0.69 and 0.7 respectively. By removing deceased patients from the training set (Alive Only) we can notice an improvement in model prediction, achieving an F1 score of 0.71 and a ROC-AUC of 0.72. It is important to note that this model is evaluated only on patients contained in the test set who were released alive, which is a relevant limitation of this method.

The approach based on WOC leads to noticeable performance improvements. In fact, the WOC model achieves 0.75 in both the F1 score and the ROC-AUC, improving the performance of the Base model by 5 points on average.

Finally, the best results are achieved by the WOC + Outcome model with a value of 0.76 in both the F1 score and the ROC-AUC. This confirms how the use of an additional feature that suggests the patient's outcome can improve performance in the SLOS task. For the Base + Outcome model, we can see how using the prediction of the outcome alone produces slightly worse results in terms of F1 score and ROC-AUC. However, there is a significant improvement in terms of F2 score. In our opinion, this is due to the conservative approach we adopted in optimising the outcome model on the F2 score, which forces the model to focus on avoiding false negatives at the cost of generating some false positives (as we showed in [13]). Moreover, the errors introduced by the Outcome Model, which increased the noise in the data, lead our best model to a slight loss in performance.

In Figure 3b we report the confusion matrices of the WOC + Outcome model. In the Alive matrix, the errors are almost equally distributed between the two classes, as confirmed by the high value of the F1 score obtained by the model. Furthermore, the model can classify patients who have a short stay with greater confidence, which are the majority. However, the Dead matrix shows a more interesting result. In fact, in our test set, only 15 of the 208 patients who died during hospitalisation were classified as short stay. This shows that our model can identify patients recovered under critical conditions and assigns them to the Long-Stay class, according to the WOC approach, leveraging the prediction made by the Outcome Model.

5.1. Feature Importance

We also analysed the contribution of the features used by our models through SHAP (SHapley Additive exPlanations), one of the most important methods for explaining the prediction of an instance that a machine learning model makes [20, 21]. This method is based on assigning to each feature a value, called *Shapley value*, which summarises

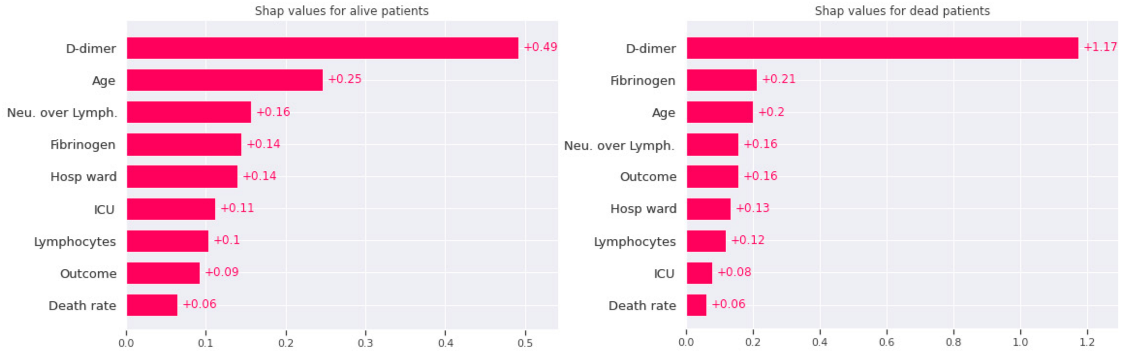


Fig. 4: Average Shapley values of the most important features in the WOC + Outcome model, calculated using the SHAP algorithm, for patients released alive (on the left) and dead (on the right). Where Neu. over Lymph. is the ratio between Neutrophils over Lymphotices.

its importance in the prediction made. The average Shapley value for all samples can be used to identify the most important features of the model in general. We calculated the Shapley values of our features for the best-performing models for each approach. Figure 4 shows the average Shapley values of the most important features used for the classification of patients who were discharged alive (left) and for the patient who died during hospitalisation (right) in our best model (WOC + Outcome). From our analysis, we can see that the D-dimer is a very important feature in determining the SLOS class. Furthermore, D-Dimer can provide an assessment of the severity of COVID-19, as shown in recent medical studies [32]. We also see that the features Age, Neutrophilis over Lymphocytes and Fibrinogen are very important for both types of prediction. An interesting result is obtained for the Outcome feature. In fact, its Shapley value highlights that this feature is more important for deceased patients than for others. This confirms the capability of the WOC approach to use the predicted outcome to better classify dead patients who will remain for more than a week.

6. Conclusions and Future Work

In this paper, we proposed a system to evaluate the conditions of COVID-19 patients to identify those patients who stay longer than a week. We trained and tested several machine learning algorithms (Random Forests, ExtraTrees, XGBoost, Feed-Forward Neural Networks) using only demographic information and laboratory tests from more than 6000 patients hospitalised by the *Spedali Civili* Hospital in northern Italy. Our patients are distributed over a long period of time, from March 2020 to December 2021, forcing us to deal with important changes such as Alpha and Delta variants, the introduction of more effective treatments, and the use of vaccines.

Experimental results show good performance in classifying patients with long and short stay, achieving a value of 0.76 in both the F1 score and ROC-AUC with our best model. In particular, we show that applying the WOC approach leads to a noticeable improvement in performance. Moreover, we show that the outcome feature, provided by integrating a model that predicts the patient's outcome in our architecture, provides a boost in the performance if applied together with the WOC. The analysis of the most important features shows that D-Dimer, Neutrophilis over Lymphocytes, and Fibrinogen are the most important laboratory tests to recognise the length of a patient's stay in the hospital. The same values are identified as important in identifying the severity of a patient in other works (e.g. [31, 32]).

Acknowledgements

This work has been partially supported by the Italian Ministry of Research and University, "Fondo Integrativo Speciale per la Ricerca (FISR)", project *AICovidPrognosis2020* (FISR grant code: FISR2020IP_04373).

References

[1] Abdulaal, A., Patel, A., Charani, E., Denny, S., Mughal, N., Moore, L., et al., 2020. Prognostic modeling of covid-19 using artificial intelligence in the united kingdom: model development and validation. Journal of Medical Internet Research 22, e20259.

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623–2631.

[3] Alafif, T., Tehame, A.M., Bajaba, S., Barnawi, A., Zia, S., 2021. Machine and deep learning towards covid-19 diagnosis and treatment: survey, challenges, and future directions. International journal of environmental research and public health 18, 1117.

[4] An, C., Lim, H., Kim, D.W., Chang, J.H., Choi, Y.J., Kim, S.W., 2020. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide korean cohort study. Scientific reports 10, 1–11.

[5] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[6] Brock, G.N., Barnes, C., Ramirez, J.A., Myers, J., 2011. How to handle mortality when investigating length of hospital stay and time to clinical stability. BMC medical research methodology 11, 1–14.

[7] Bullock, J., Luccioni, A., Pham, K.H., Lam, C.S.N., Luengo-Oroz, M., 2020. Mapping the landscape of artificial intelligence applications against covid-19. Journal of Artificial Intelligence Research 69, 807–845.

[8] Chen, J., Li, K., Zhang, Z., Li, K., Yu, P.S., 2021. A survey on applications of artificial intelligence in fighting against covid-19. ACM Computing Surveys (CSUR) 54, 1–32.

[9] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

[10] Chiari, M., Gerevini, A.E., Maroldi, R., Olivato, M., Putelli, L., Serina, I., 2021a. Length of stay prediction for northern italy covid-19 patients based on lab tests and x-ray data, in: International Conference on Pattern Recognition, Springer. pp. 212–226.

[11] Chiari, M., Gerevini, A.E., Olivato, M., Putelli, L., Rossetti, N., Serina, I., 2021b. An application of recurrent neural networks for estimating the prognosis of covid-19 patients in northern italy, in: International Conference on Artificial Intelligence in Medicine, Springer. pp. 318–328.

[12] Daneshkohan, A., Alimoradi, M., Ahmadi, M., Alipour, J., 2022. Data quality and data use in primary health care: A case study from iran. Informatics in Medicine Unlocked , 100855.

[13] Gerevini, A., Maroldi, R., Olivato, M., Putelli, L., Serina, I., 2020. Prognosis prediction in covid-19 patients from lab tests and x-ray data through randomized decision trees, in: 5th International Workshop on Knowledge Discovery in Healthcare Data, KDH 2020, pp. 27–34.

[14] Gerevini, A.E., Lavelli, A., Maffi, A., Maroldi, R., Minard, A., Serina, I., Squassina, G., 2018. Automatic classification of radiological reports for clinical care. Artificial Intelligence in Medicine 91, 72–81.

[15] Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine learning 63, 3–42.

[16] Johnson, A.E., Ghassemi, M.M., Nemati, S., Niehaus, K.E., Clifton, D.A., Clifford, G.D., 2016. Machine learning and decision support in critical care. Proceedings of the IEEE 104, 444–466.

[17] Kawahara, J., Hamarneh, G., 2018. Fully convolutional neural networks to detect clinical dermoscopic features. IEEE journal of biomedical and health informatics 23, 578–585.

[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30.

[19] Lam, S.W.S., Abdullah, H.R.B., Pourghaderi, A.R., Nguyen, N.H.L., Wu, J.T., Dev, S., Mohan, O., Low, S.K., Lee, J.K., Tan, B.R., et al., 2020. Towards health system resiliency: An agile systems modelling framework for bed resource planning during covid-19. BMC pre-print .

[20] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable ai for trees. Nature machine intelligence 2, 56–67.

[21] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777.

[22] Nemati, M., Ansary, J., Nemati, N., 2020. Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data. Patterns 1, 100074.

[23] Putelli, L., Gerevini, A.E., Lavelli, A., Maroldi, R., Serina, I., 2021. Attention-based explanation in a deep learning model for classifying radiology reports, in: Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings, Springer. pp. 367–372.

[24] Putelli, L., Gerevini, A.E., Lavelli, A., Olivato, M., Serina, I., 2020. Deep learning for classification of radiology reports with a hierarchical schema, in: Cristani, M., Toro, C., Zanni-Merk, C., Howlett, R.J., Jain, L.C. (Eds.), Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES-2020, Virtual Event, 16-18 September 2020, Elsevier. pp. 349–359.

[25] Putelli, L., Gerevini, A.E., Lavelli, A., Serina, I., 2019. Applying self-interaction attention for extracting drug-drug interactions, in: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings.

[26] Sharma, S., Gupta, Y.K., 2021. Predictive analysis and survey of covid-19 using machine learning and big data. Journal of Interdisciplinary Mathematics 24, 175–195.

[27] Subbaswamy, A., Saria, S., 2020. From development to deployment: dataset shift, causality, and shift-stable models in health ai. Biostatistics 21, 345–352.

[28] Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. Journal of machine learning research 1, 211–244.

[29] Vekaria, B., Overton, C., Wiśniowski, A., Ahmad, S., Aparicio-Castro, A., Curran-Sebastian, J., Eddleston, J., Hanley, N.A., House, T., Kim, J., et al., 2021. Hospital length of stay for covid-19 patients: Data-driven methods for forward planning. BMC Infectious Diseases 21, 1–15.

[30] Yadaw, A.S., Li, Y.c., Bose, S., Iyengar, R., Bunyavanich, S., Pandey, G., 2020. Clinical features of covid-19 mortality: development and validation of a clinical prediction model. The Lancet Digital Health 2, e516–e525.

[31] Yan, L., Zhang, H.T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., et al., 2020. An interpretable mortality prediction model for covid-19 patients. Nature Machine Intelligence 2, 1–6.

[32] Yao, Y., Cao, J., Wang, Q., Shi, Q., Liu, K., Luo, Z., Chen, X., Chen, S., Yu, K., Huang, Z., et al., 2020. D-dimer as a biomarker for disease severity and mortality in COVID-19 patients: a case control study. Journal of intensive care 8, 1–11.

[33] Žliobaitė, I., Pechenizkiy, M., Gama, J., 2016. An overview of concept drift applications. Big data analysis: new algorithms for a new society , 91–114.