Medical & Biological Engineering & Computing (2022) 60:459–470
https://doi.org/10.1007/s11517-021-02479-8

ORIGINAL ARTICLE

Predicting SARS-CoV-2 infection duration at hospital admission:a deep learning solution

Piergiuseppe Liuzzi[1,2] · Silvia Campagnini[1,2] · Chiara Fanciullacci[2] · Chiara Arienti[3] · Michele Patrini[3] · Maria Chiara Carrozza[1] · Andrea Mannini[1,2]

Received: 2 April 2021 / Accepted: 24 November 2021 / Published online: 7 January 2022

Abstract
COVID-19 cases are increasing around the globe with almost 5 million of deaths. We propose here a deep learning model capable of predicting the duration of the infection by means of information available at hospital admission. A total of 222 patients were enrolled in our observational study. Anagraphical and anamnestic data, COVID-19 signs and symptoms, COVID-19 therapy, hematochemical test results, and prior therapies administered to patients are used as predictors. A set of 55 features, all of which can be taken in the first hours of the patient's hospitalization, was considered. Different solutions were compared achieving the best performance with a sequential convolutional neural network-based model merged in an ensemble with two different meta-learners linked in cascade. We obtained a median absolute error of 2.7 days (IQR = 3.0) in predicting the duration of the infection; the error was equally distributed in the infection duration range. This tool could preemptively give an outlook of the COVID-19 patients' expected path and the associated hospitalization effort. The proposed solution could be viable in tackling the huge burden and the logistics complexity of hospitals or rehabilitation centers during the pandemic waves.

Keywords Artificial intelligence · Convolutional neural network · COVID-19 · Duration of infection · Prognostic models · Rehabilitation

1 Introduction

Since October 2020, almost 300 million people have been infected by SARS-CoV-2 with more than 5 million of deaths (World Health Organization, WHO, reports). It is well known that in severe cases, treatment in intensive care units is required. This can lead to overcrowding of hospitals and rehabilitation settings [1] that is currently posing a global burden to healthcare systems [2, 3]. As already investigated for other pathological conditions [4, 5], artificial intelligence (AI) is being applied to extract predictive information with the potential to revolutionize the approach to tackle COVID-19 [6, 7]. Prediction models, capable of correlating patients' characteristics to the evolution traits of the disease and possible patients' responses to it, can provide helpful support to the decision-making process in clinical environments [8–11].

For what concerns COVID-19 prognostic models, literature mainly focuses on mortality risk, assessing it at admission [11], after a week [12], or predicting the discharge setting [13]. To the extent of our knowledge, only two articles attempted to find a relation between length of stay and predictive features [14, 15]. Wang et al. [14] showed that patients in high risk and low risk (identified by using the features with most predictive power in their diagnostic model) had significant difference in length of stay. Qi et al. [15] instead targeted short-term hospital stay (< 10 days) and long-term hospital stay (> 10 days) and obtained a binary classification.

In many pathological contexts, the length of stay is often addressed as an outcome, considered both as an indirect index of severity of the disease and an essential data for hospitals administration. However, especially during pandemic

✉ Silvia Campagnini
scampagnini@dongnocchi.it

[1] Scuola Superiore Sant'Anna, The BioRobotics Institute, Viale Rinaldo Piaggio 34, 56025 Pontedera, PI, Italy

[2] IRCCS Fondazione Don Carlo Gnocchi, via di Scandicci 269, 50143 Firenze, FI, Italy

[3] IRCCS Fondazione Don Carlo Gnocchi, via Alfonso Capecelatro 66, 20148 Milano, FI, Italy

outbreaks, length of stay appears to be highly impacted by external factors like personnel/bed availability and local differences in hospital management rules. Focusing on infection duration looks as a promising solution in these regards, thanks to its capability to overcome length of stay limitations and keep at the same time the aforementioned duality trait. Up to our knowledge and to updated systematic reviews [16], no existing research assesses the specific problem of infection duration by means of data-driven regression models. This knowledge in our view could lead to an innovative tool to be implemented in electronic health record (EHR) allowing for a significant advantage in the management of both clinical and administrative aspects for COVID-19 patients' treatment. Indeed, it could ease the practitioner in the delivery of personalized care to patients as well as supporting management of beds, intensive care units, and ventilation units.

To fill this gap, starting from a dataset of 222 COVID-19 patients treated in the Fondazione Don Gnocchi hospital network, we compared different machine learning solutions with the aim of predicting the duration of the infection. The resulting most performant solution was based on a convolutional neural network (*CNN*) model (namely, CNN-*core*) and it was obtained by four steps: (1) training of the CNN-*core*, (2) combining the cores in an ensemble, (3) adding two separate meta-learners (logistic regression and fully connected neural network), and lastly, (4) voting among meta-learners predictions. The achieved accuracy (median infection duration absolute error of 2.7 days, IQR = 3.0 days) looks promising for the implementation of a decision support tool to be integrated with the EHR of the hospital network.

2 Methods

2.1 Study design and participants

An observational study was performed including 518 patients who were discharged from 16 Fondazione Don Gnocchi centers involved in the COVID-19 patients' care. Inclusion criteria were based on current or previous infection by SARS-CoV-2 virus at admission in hospital and thus all patients positive to COVID-19 (age $\geq$ 18 years) were enrolled in the study. All patients were diagnosed with COVID-19 strictly following WHO guidelines [17]. Positive cases were verified maximally every 10 days via molecular tests. Due to the high spectrum of cases, patients in the database were primarily classified into as follows: type 1, already positive to SARS-CoV-2 before admission; type 2, turned positive during their stay; type 3, hospitalized after the infection for rehabilitation purposes. Given that the target of this study was the estimation of the duration of the infection from hospital admission data, only type 1 patients were retained for further analyses (222 patients). These data referred to the first pandemic wave in Italy and were retrospectively acquired from April to September 2020.

Especially during the first pandemic wave, the emergency scenario and the lack of treatment protocols for an unknown disease played a role in increasing the heterogeneity among patients' characteristics. For instance, an aspect of interest for our study was the time difference between the admission to the IRU and the first negative test with no subsequent positive ones (median 12 days, IQR = 20.5). These numbers gave us a further confirmation in targeting the infection duration as outcome, considering it as a more reliable and less regional-dependent proxy of hospitalization than the length of hospital stay. The infection duration, measured in days, was calculated as the difference between the date of the first positive molecular test and the date of the following first negative one, without subsequent positives. The infection duration was finally calculated when at least two negative results were collected. Indeed, this variable hosts more general information with respect to the length of stay and that it can be more versatile. In fact, it can be applied independently of the differences in healthcare organizations in different regions/countries and independently from the specific emergency status of the healthcare system at the time of the recovery.

2.2 Data collection, preprocessing, and preliminary statistical analysis

The study protocol has been approved by the Ethical Committee of the IRCSS Fondazione Don Carlo Gnocchi the 16/04/2020.

A structured data collection was designed on REDCap (Research Electronic Data Capture, Vanderbilt University, 2021 West End Avenue, Nashville, TN 37,235, USA), an online-based software for database development. The database was structured in a way to collect each evaluation or assessment in four distinct events: admission, during the recovery, discharge, and, only in the case of type 3 subjects, the acute phase of the disease. However, for specific data groups, such as the results of the diagnostic tools, it was given the disposition to collect information about any test repeated, independently on the events planned.

More than 800 features have been taken and the complete dataset includes anagraphical data, symptoms and vital signs, hematochemical and hemogasanalysis values, instrumental data (RX, CT, EEG, etc.…), multiple assessments (cognitive, psychological, functional, and nutritional), and prior clinical data [18].

The median age of the 222 patients included in the study was 76 (IQR = 19) and the male was the 46% of the total dataset. The median infection duration was 31 days (IQR = 26) and the values fell in a range between 11 and

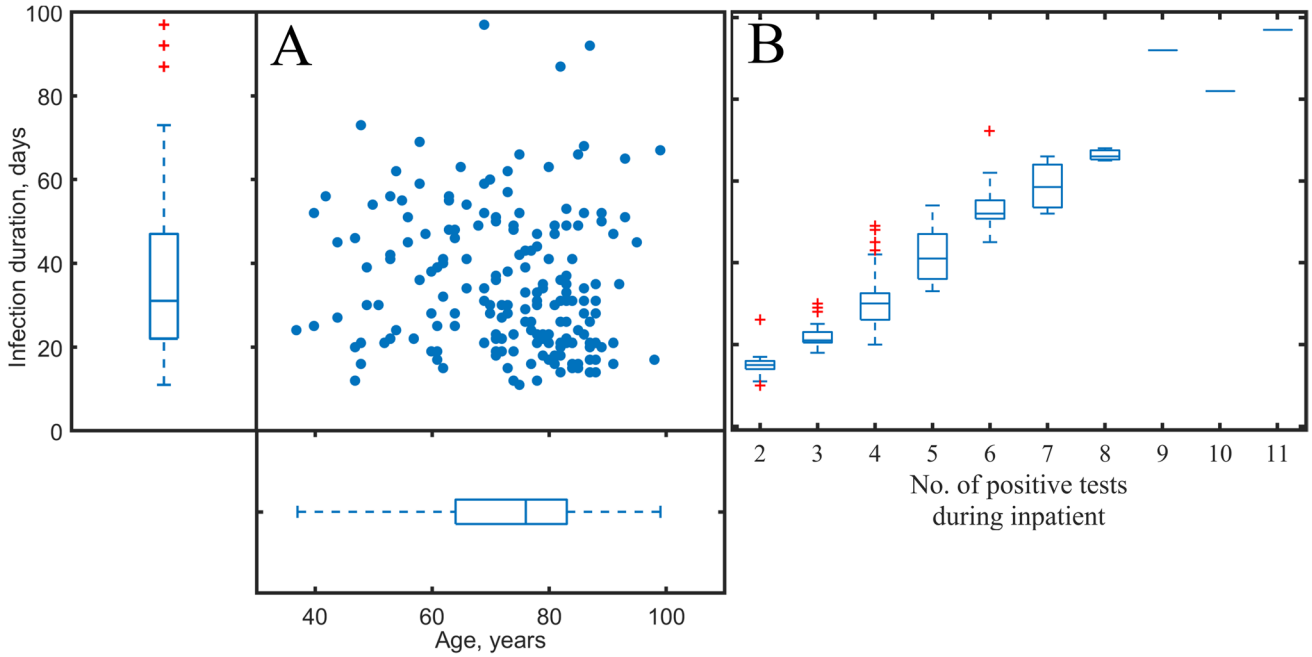


Fig. 1 Distribution of the age with respect to infection duration and their respective box-plots

97 days (Fig. 1, panel A). An almost linear relationship between infection duration and number of molecular tests can be observed (Fig. 1, panel B).

Preliminary statistical analyses were carried out in SPSS (Vs 26, Chicago, SPSS Inc.). They concerned univariate analysis to understand the influence of each selected predictor with respect to the infection duration. In particular, Pearson correlations were applied with numerical variables, while the non-parametric Mann–Whitney test was applied for dichotomous variables. The data preprocessing was conducted in Matlab (R2019b, The MathWorks, Inc., Natick, MA, USA) as well as the machine learning (ML) models. The deep learning (DL) models were written in Python 3.0 (Python Software Foundation) using the TensorFlow library. Pseudonymized data can be made available upon request to researchers to validate and reproduce results.

For the comparison of different machine learning methods, we presented as touchstone the linear regression, a simple and interpretable model. To compare the different machine learning solutions proposed, median absolute errors of each solution were compared with the linear regression error on the same population by mean of Wilcoxon signed-rank tests. Moreover, an effect size of this comparison was calculated as the ratio of negative differences between the measurements and the total numerosity.

2.3 Feature screening and dimensionality reduction

As it was already pointed out, the initial dataset was composed by 829 features. Firstly, its dimensionality was reduced by keeping features taken in the first 8 h from admission and the ones with a fill percentage higher than the 30% of the column length (total subjects). Secondly, via a literature search of relevant correlates, we included supports, sign and symptoms, and clinical and hematochemical data. Furthermore, pharmacological therapies (both COVID related and non-COVID related) were included in the feature set given their availability in the dataset at the time of admission.

Missing data in the training, validation, and test sets were substituted by the mean (for numerical data) or the mode (for the categorical data) of the correspondent variable in the training set, reaching a full set of 55 features (Table 1).

A further reduction on data dimensionality was then achieved by principal component analysis (PCA). Five principal components were retained yielding a variance > 99%. The same PCA transform was then applied to the test set.

2.4 Model architecture

In order to test different approaches to the problem, ML (linear regressions, random forests) and DL (convolutional neural network) models were compared. We tackled the problem with an approach of growing complexity. Regularized linear regression and random forests were considered because of the simpler interpretability of the model, which is a non-negligible aspect in the clinical practice. More complex architectures (CNN) were subsequently developed to increase accuracy and reliability of the tool.

Table 1 Predictors used in the final model, before entering the PCA analysis. [a]Cumulative Illness Rating Scale (CIRS), [b]COVID-19 therapy prescribed prior to admission. Numerical features are italicized while categorical features are reported with regular font

Data group	Feature name	Median and IQR (numerical) or relative positive frequency (binary)
Anagraphical data (3)	*Age [years]*	*75.5 [IQR = 22]*
	Sex [1 female]	45.79%
	RSA [1 if patients comes from residential care unit]	0.52%
Admission clinical scales (3)	*ICD (number of events)*	*3 [IQR = 2]*
	CIRS severity index[a]	*1.4 [IQR = 0.4]*
	CIRS comorbidity index	*2 [IQR = 2]*
Admission signs and symptoms (2)	Fever	58.42%
	Dyspnea	44.74%
Admission supports (8)	Invasive mechanical ventilation (IMV)	36.84%
	O_2 therapy	56.32%
	IMV or O_2 therapy	62.22%
	Extracorporeal membrane oxygenation (ECMO)	1.58%
	Urinary catheter	42.63%
	Tracheal cannulation	13.68%
	Artificial alimentation	14.21%
	Venous cannulation	33.68%
COVID-19 therapy[b] (17)	Favipiravir	8.42%
	Avigan	7.59%
	Tocilizumab	2.11%
	Remdesivir	37.89%
	Lopinavir-ritonavir association	20%
	Darunavir	65.79%
	Cobicistat	65.79%
	Ruxolitinib	0.52%
	Ribavirin	1.05%
	Hydroxychloroquine	40.52%
	Azithromycin	0.52%
	Colchicine	3.16%
	Heparin	66.32%
	Enoxaparin sodium	0.52%
	Baricitinib	36.84%
	Corticosteroids	62.11%
	Other antibiotics different from azithromycin	27.90%

Table 1 (continued)

Data group	Feature name	Median and IQR (numerical) or relative positive frequency (binary)
Therapy prior to COVID-19 (17)	ACE inhibitors	11.05%
	Sartans	7.37%
	Antimineralocorticoid	11.58%
	Antiplatelet	19.47%
	Anticoagulant	35.79%
	Statin	26.84%
	Beta-blockers	26.32%
	Calcium channel blockers	1.05%
	Amiodarone	1.05%
	Non-steroidal anti-inflammatory drug	3.16%
	Steroid therapy	1.58%
	Levodopa	18.42%
	Immunosuppression	0%
	Anxiolytic-antidepressant	33.68%
	Proton-pump inhibitor	6.84%
	Vitamines	3.68%
	Other therapies	3.16%
Hematochemics (5)	*White blood cells [n. samples* $\times 10^9$*/l]*	*6.4 [IQR = 3.26]*
	Neutrophils [n. samples $\times 10^9$*/l]*	*4.2 [IQR = 2.43]*
	Lymphocytes [n. samples $\times 10^9$*/l]*	*1.46 [IQR = 0.90]*
	Hemoglobin [g/l]	*97.5 [IQR = 106.45]*
	Platelets [n. samples $\times 10^9$*/l]*	*292 [IQR = 158]*

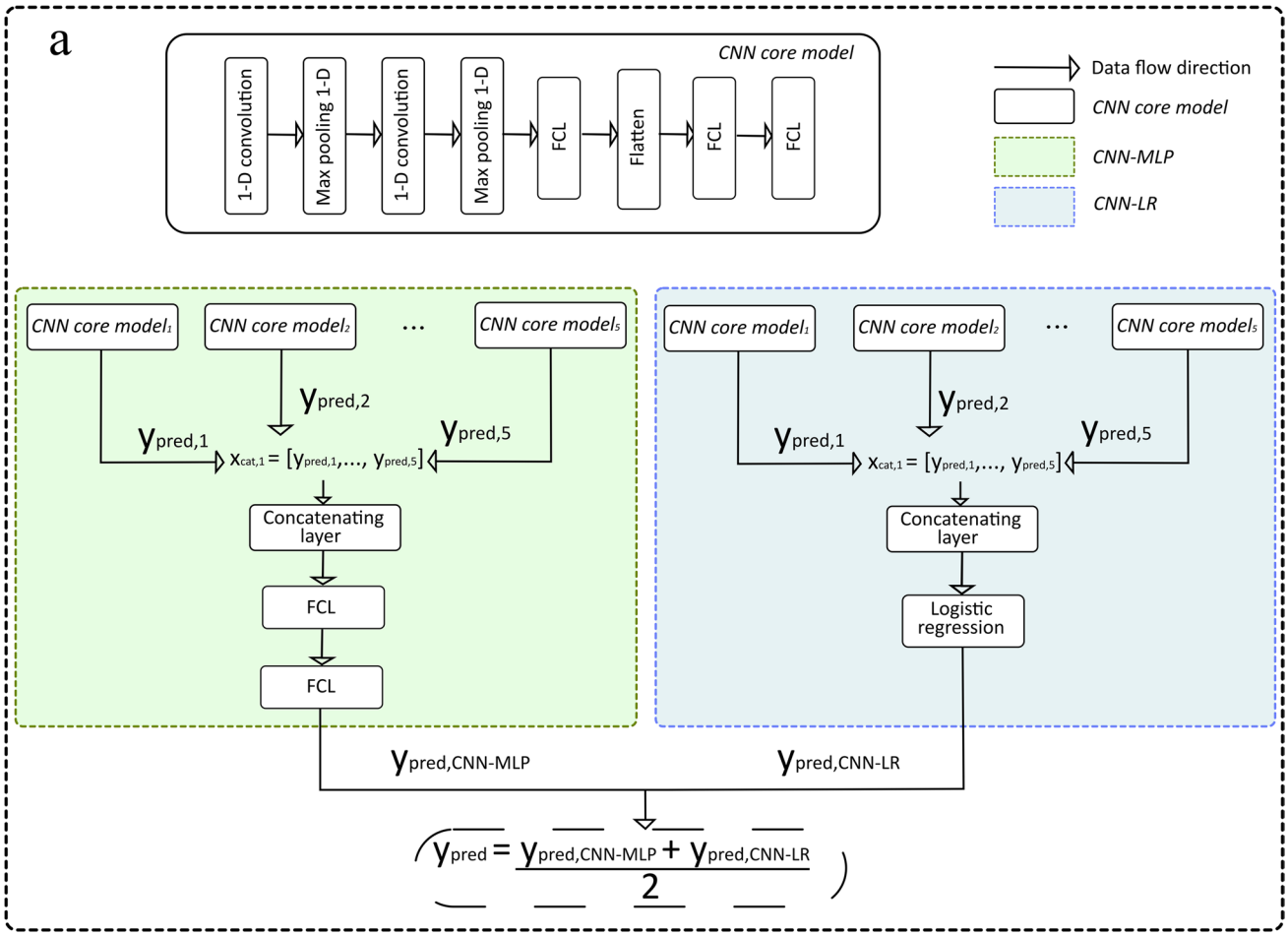
The random forest model is an ensemble learning method using a regression tree as template learner [19, 20]. In such model, a set of binary decision trees are merged in a single ensemble classifier and the input features of each tree are subsamples of the available features. To define the model, both the minimum number of leaf node observations in each tree and the number of predictors to sample (n_{PTS}) at each node need to be selected.

CNNs are a type of neural network capable of hierarchically assembling more structured patterns using simpler ones [21]. The CNN-*core model* we choose was composed by sequential layers of different type (Fig. 2a). Together with the 1-D convolutional layers, 1-D max pooling layers were implemented. The aim was to reduce data dimensionality by combining the outputs of neuron clusters at the prior layer into a single neuron in the subsequent layer. It has been previously demonstrated that such a convolution-pooling, fully connected structure, can successfully process both images [22] and one-dimensional data [23, 24].

2.5 Training and testing

Each model was trained using as input the PCA-transformed data and the infection duration as target. A train-validation-test split was done to validate the model over a different number of subjects and parameters. The validation strategy adopted was *K*-fold validation with $N_{folds} = 5$. The test portion removed corresponds to 15% of the samples (33 subjects); hence, each fold of the cross-validation was composed by 37 or 38 subjects.

Additionally, the whole process was repeated 10 times: the obtained aggregated results were reported to reduce the effect of randomized parameter initialization (CNN) and randomized test split.



Training pseudo-code

```
# Hyperparameter optimization
opt_par = gridSearchCV (xtr, ytr)
# CNN core model building and training
for k = 1:5
        mdl[k] = build_CNNcoremodel (opt_par)
        mdl[k] = trainCNNcoremodel (mdl[k], xtr, ytr)
        tr_err[k] = getTrainError (mdl[k])
end
ensemble_model = concatCNN(mdl)        # Horizontally stack models
# CNN-MLP building and training (CNN core model weights are retrained)
MLPmodel = build_CNN_MLP (ensemble_model)
MLPmodel = train_CNN_MLP (MLPmodel, xtr, ytr)
# CNN-LR building and training (CNN core model weights are not retrained)
LRmodel = build_CNN_LR (ensemble_model)
LRmodel = train_CNN_LR (LRmodel, xtr, ytr)
```

b

Testing pseudo-code

```
# CNN core model testing
Ypred,CNNcore = getPredictions(mdl[0], xtest[i])
# CNN-MLP model testing
Ypred,CNNcore,all = []
for k = 1:5
        Ypred,CNNcore,k = getPredictions(mdl[k], xtest[i])
        if detrending
                # Subtract median train error from prediction
                Ypred,CNNcore,k = Ypred,CNNcore,k – tr_err[k]
        end
        # Stack predictions horizontally
        Ypred,CNNcore,all = [Ypred,CNNcore,all, Ypred,CNNcore,k]
end
# Provide inputs to stage 2 and get CNN-MLP prediction
Ypred,CNN-MLP = getPredictions(MLPmodel, Ypred,CNNcore,all)

# CNN-LR model testing with same procedure
Ypred = (Ypred,CNN-MLP + Ypred,CNN-LR)/2   # Voting
```

c

◂**Fig. 2** Panel (a): CNN-core model and its integration in CNN-MLP and CNN-LR ensembles involving MLP and LR meta-learners (green and blue boxes respectively). Panel (b): pseudo-code for the training of the metalearners. Panel (c): pseudo-code of the validation and testing phase

2.6 Hyper-parameter optimization

In the ML models, the regularization parameter λ for the linear regression, the complexity (depth) of the trees, and the number of predictors to sample at each node for the random forest were optimized by grid search.

For the CNN, given the large number of hyper-parameters to be chosen, a more complex grid search was conducted.

The involved parameters were the adaptive moment estimation (Adam) optimizer learning rate l_r, the number of neurons in the first and second fully connected layer $n_{\text{neurons},1}$ and $n_{\text{neurons},2}$, the number filters in the two convolutional layers $n_{\text{filters},1}$ and $n_{\text{filters},2}$, and the number of training epochs n_{epochs}. The neuron activation functions were chosen to be Rectified Linear Units (ReLUs) with $f_{\text{ReLU}}(a) = \max(0, a)$. The third and last fully connected layer (FCL) uses a linear activation function $f_{\text{Lin}}(a) = a$.

The range of each variable is shown in Table 2 and for each permutation ($N_{\text{perm}} = 4 \times 5 \times 4 \times 3 \times 3 \times 4 = 2880$ different model configurations), the optimization process was run 5 times per configuration and the result aggregated. The final configuration was chosen as the one of the model with minimum validation error.

2.7 Ensembling

In order to improve the performance of the CNN-*core* model, a stacked ensemble learning approach was implemented. It was done by concatenating the individual *CNN-core* models ($N_{CoreCNN} = 5$) predictions into a second feature vector (Fig. 2a). The individual *CNN-core* models differed only for starting weight initialization and a different random number generator seed in the Adam stochastic optimization. $X_{cat,1}$ will be fed to the learning stage 2, also called meta-learner, in order to reduce the inductive training bias and the effect of random weight initialization of the single sub-models on test predictions.

As meta-learner, two different multi-layer perceptron were implemented: (i) a logistic regression (CNN-LR) and (ii) a fully connected neural network (CNN-MLP).

The meta-learner (stage 2) training followed the training of each of the models in the learning stage 1. For the logistic regression (LR), the learning stage 1 weights were kept constant to the final weights of the respective training phase during training of the meta-learner, since no back-propagation training is required for logistic regressions. Conversely, while training the MLP, the learning stage 1 weights were re-trained with starting weights set equal to the final weights of their previous training phase (Fig. 2b). The reason behind this is that the logistic regression does not need a back-propagation training while the MLP does.

2.8 Voting

Unlike meta-learning, during a voting process, each model output is considered with the same weight. In regression tasks, we can increase the performance of the overall model, balancing the offsets of the single sub-models, by averaging among their predictions [25].

Hence, CNN-MLP and CNN-LR predictions were averaged to obtain the final result via $y_{\text{pred}} = \frac{y_{\text{pred,CNN-MLP}} + y_{\text{pred,CNN-LR}}}{2}$. Furthermore, to improve the re-training and the voting process, hence reduce each of the *CNN-core* models bias, we removed from each of the 5 core test predictions the median prediction error of its training set. Then, it was fed again to the meta-learning stage 2. This resulted in an improved approach allowing the subsequent procedures (ensembling, meta-learning, and voting) to yield more accurate estimates.

3 Results

From the preliminary biostatistical analyses, Pearson correlations with the infection duration were found to be statistically significant for the Cumulative Illness Rating Scale (CIRS) [26] declined as severity and comorbidity indexes (p-values respectively of 0.001 and 0.003) (Table 1, Fig. 3).

For what concerns features related to therapies, patients with an ongoing therapy with tocilizumab ($p = 0.033$), vitamins ($p = 0.02$), anticoagulants ($p = 0.044$), calcium channel blockers ($p = 0.019$), and anxiolytic-antidepressant (weak, $p = 0.054$) showed a statistically significant longer duration of the infection. Finally, concerning vital support aids, only the presence of the tracheal cannula showed weak association with a p-value of 0.053 (Fig. 3).

For what concerns the automatic prediction of the outcome, after optimizing hyper-parameters for all the tested methodologies (Table 2), the linear regression resulted in a median absolute error of 13.23 days (IQR = 10.19), while the random forest, with 15.39 days (IQR = 13.95), performed slightly worst (Fig. 4). The grid-search for the *CNN-core* hyper-parameter optimization resulted in the best configuration having train, validation (fivefold), and test median AE of 11.12, 11.35, and 9.63 days respectively.

After combining 5 *CNN-core* models in a stacked ensemble, adding two meta-learners (LR and MLP) and voting among the two models, the median test error resulted to be 4.67 days (IQR = 5.25). However, it can be noticed that the predictions in this case resulted skewed from the ideal output (Fig. 5, orange markers). Detrending the *CNN-core* test

Table 2 Grid values for the optimization of the ridge linear regression (A), random forest (B), and convolutional NN (C). Subscripts refer respectively to the FCL layers (for the number of neurons) and to the convolutional layers (for the number of filters). The output FCL layer is a single-output neuron, being this a single-output regression

Optimized variable	Search range	Best A
λ	[0–10] with step 0.1	1.1
Optimized variable	**Search range**	**Best B**
n_{PTS}	[1,5,50,100]	5
LS_{min}	[1,5,10,15,20]	20
Optimized variable	**Search range**	**Best C**
l_r	[0.00001, 0.0001,0.01,0.1]	0.001
$n_{\mathrm{neurons},1}$	[32,64,128,256,512]	256
$n_{\mathrm{neurons},2}$	[8,32,64,128]	128
$n_{\mathrm{filters},1}$	[16,32,64]	32
$n_{\mathrm{filters},2}$	[12,32,64]	64
n_{epochs}	[5,10,20,50]	5

predictions by removing the respective train error before the ensembling and voting processes solved the problem leading to a median test error of 2.5 days (IQR = 1.92).

The solution based on "CNN ensemble + voting" and the one including "CNN detrended + ensemble + voting" showed significantly improved accuracies with respect to the linear regression, as confirmed by the Wilcoxon signed rank test (Fig. 5). In the comparison between the random forest and the *CNN-core* model, no significant differences in performance were obtained. Moreover, the effect sizes of the CNN model after voting (0.78) and of the detrended CNN model after voting (0.94) statistically confirm the improvement of our model with respect to the linear regression ($p < 0.001$).

The aggregated results after the repetition of the procedure with the same hyper-parameters multiple times ($N_{\mathrm{run}} = 10$) are summarized in Fig. 5. The median error of 2.7 days is very similar to the one obtained with only one run (2.5), but the IQR is higher (3.0 days for 10 runs compared to 1.9 for 1 run). The determination coefficient (R^2), calculated between real and predicted values, was positively impacted by both ensembling/voting and detrending procedures, reaching $R^2 = 0.91$ for the final solution.

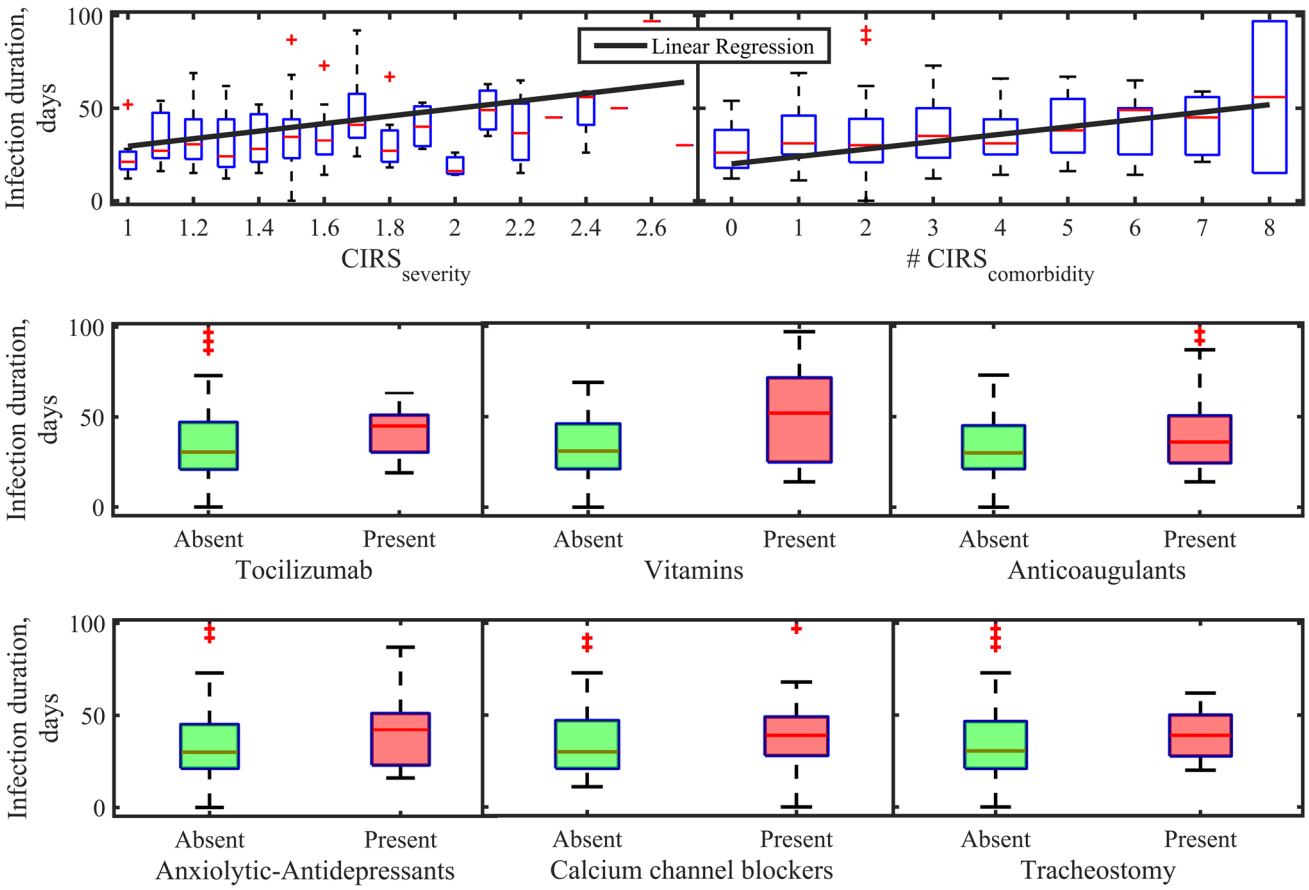


Fig. 3 Graphical representation of the infection duration versus significant variables. Results from the CIRS severity and comorbidity index correlation with infection duration are reported in the upper panels while group comparisons are reported in the middle and lower panels

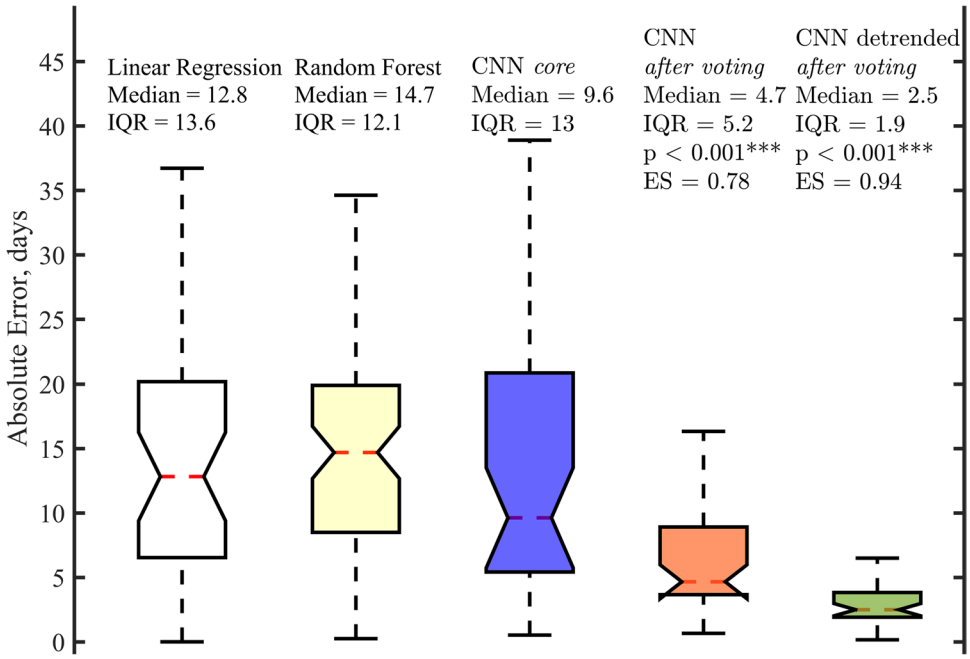


Fig. 4 Box plot of the absolute error (days). The CNN result is referred to the CNN-core, while the CNN after voting refers to the model after being combined in the ensemble/meta learning step and the voting procedure (with no detrending). The CNN detrended + voting plot corresponds to the final result with detrending, preceding ensembling, and voting (absolute test error calculated on y_{pred} after voting between CNN-MLP and CNN-LR)

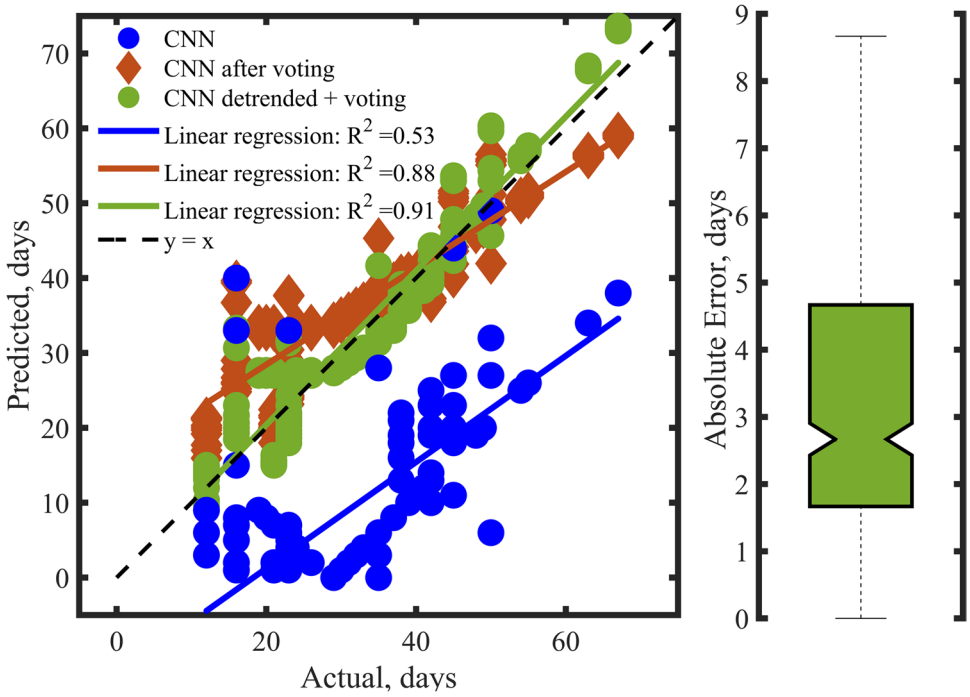


Fig. 5 Left panel: Scatter plot of predicted infection duration with respect to the real value. Values obtained by aggregating together the output of 10 different runs of the procedure with fixed hyper-parameters. Results obtained with three different methods. Right panel: Absolute prediction error (calculated from the 10 different run aggregated together) of the best performing solution (green dots in the left panel)

4 Discussion

In this study, a predictive model for the duration of SARS-CoV-2 infection in hospitalized patients was investigated and validated on data from 222 patients. Classical machine learning algorithms, such as optimized linear regressions and random forests, resulted in performances not fully satisfying for this problem. However, non-linear models resulted to significantly improve the prediction accuracy. Indeed, on our dataset, a model of increased complexity is needed for an accurate prediction of the clinical outcome at the expense of a reduced interpretability. Our cross-validation results confirm that, by means of data taken in the first 8 h from patients' admission, an accurate prediction

(median error < 3 days) of the duration of the infection is feasible.

This resulted to be a significant improvement with respect to the linear regression. Furthermore, we presented results of the steps through the development of the CNN based final model, confirming that an increased model complexity can be reached with simple techniques, as ensembling and voting.

During a time in which a complex pandemic seems still to affect importantly healthcare services, a prognostic prediction tool can support clinical decision in hospitals or sanitary structures by providing data-driven elements for a better time planning and hospital organization [27–29].

As already stated, we focused on predicting the infection duration. Up to our knowledge, there is not a previous study involving data-driven regression models targeting the infection duration for COVID-19 or any other illness. Some similar solutions reported in literature concern the length of stay estimation (Table 3). Nemati et al. [30], by means of survival analysis, targeted the in-hospital length of stay for COVID-19 patients, showing how the discharge probability reaches 1 after ~ 27 days.

Qi et al. [15], instead, focused on binary outputs as short- and long-term hospital stay (area under the curve, AUC = 0.97). By translating our best-performing regression solution in a similar binary classifier using the target median (31 days) as the threshold, we achieved an AUC of 0.98. Ebinger et al. [31] similarly classify patients according to a LoS threshold set equal to 8 days obtaining an AUC = 0.819. Lastly, Chiari et al., starting from more than 1000 patients and multimodal sources (blood exams and clinical variables), obtained a mean absolute error of 4.11 days in predicting LoS [32]. The latter manuscript presents an internally validated model, trained using a dataset with median LoS of 14 days, using data acquired up to the first 8 days after admission. Finally, Setti et al. developed a linear kernel-based support vector for regression targeting post-COVID rehabilitation LoS [27]. The model, trained on data from the first pandemic wave, was tested with data from the second pandemic wave achieving a median absolute prediction error of ~ 7 days. Regression models targeting length of stay in specific wards (MAE ~ 1 days, range: 2–7 days [33]) and in emergency unit (RMSE = 13.35 days [34]) show the complexity this prediction, by means of regression methods. Even if the comparison is not entirely fair, since our patient spectrum is narrower (only COVID-19 with respect to the heterogeneity of patients in emergency unit), we achieved a significant decrease in the prediction error.

Some relevant limitations to our work need a further discussion. In addition to the low interpretability of the model, another limit is that the infection duration could be altered by the advent of new therapies and treatments or by the diffusion of SARS-CoV-2 variants. As soon as such information will be available, a redefinition of the solutions will be necessary.

Another limitation is that our dataset was acquired in hospitals, involving symptomatic patients only. In this regard, given the simple nature of input features, it is reasonable to assume that by extending the pool of available data to the overall population, a general solution could be achieved. Still, the cohort heterogeneity for what concerns the duration of infection (from ~ 10 days up to ~ 80 days) is a point in favor of the generalizability of the results that could be improved by further patients' stratification on a larger

Table 3 Summary of literature findings on predicting COVID-19 length of stay compared with our solution

	Training and validation (# patients)	Testing (# patients)	Outcome	Results
Nemati et al. [30]	1182	–	Discharge-time probability (survival analysis)	Discharge probability = 1 after ~ 27 days. C-index from Stagewise GB = 71.47%
Qi et al. [15]	31	–	Short- and long-term hospital stay (≤ 10 days)	Data at admission, AUC = 0.97 (95% CI 0.83–1)
Ebinger et al. [31]	772	193	Short- and long-term hospital stay (≤ 8 days)	Models trained on hospital day 1–2-3. Increasing accuracies over time with an accuracy of 0.765 (AUC = 0.819) if trained on day 3
Chiari et al. [32]	524	132	Length of stay	Models trained on hospital day 2–4-6–8. Best results trained after 8 hospitalization days with a mean absolute error of 4.11 days
Setti et al. [27]	62	25	Length of stay, post-COVID rehabilitation	Data taken in the first week from admission to rehabilitation, median test error of 7.04 days [IQR = 10.7]
This study	*189*	*33*	*Infection duration*	*Data taken in the first week from admission resulted in a test median absolute error of 2.7 days [IQR = 3.0]*

database. Lastly, we have to acknowledge that a slight underestimation of the number of intermediate positive molecular tests performed on each patient may have been possible. This was primarily due to two concurrent factors, the first being the excessive burden on the healthcare facilities of the first pandemic wave and the second resulting from the fact that inferences on intermediate testing were not the main objective of the study.

Indeed, a strength of this model is that it is developed on very simple and accessible data, mostly available in the clinical routine and easily collectable in a digital form. The integration of such a model into the clinical workflow can be straightforward, through a simple graphical user interface. Even non-medical personnel can transfer the requested data into the tool, right after the admission, and obtain an estimate of the duration of the infection. This allows us to consider our tool to be "low cost" for the hospital, having at the same time an accuracy level in the estimation of infection duration which is clinically relevant.

5 Conclusions

In conclusion, we reported the development and validation of a predictive model based on data collected from Fondazione Don Gnocchi centers (Italy) during the first COVID-19 pandemic wave. This work confirms that deep learning and machine learning can be viable tools for predicting clinical outcome in order to support the clinical decision-making processes. Given the simple measurement of the input data, the model results to be easily translatable into clinical practice.

Further work will aim to perform an external prospective validation and to perform a sensitivity analysis of the prediction with respect to COVID-19 therapies and SARS-CoV-2 variants. To bring the finding of the study into clinical practice, a user-friendly software is currently under development for future integration in the clinical daily practice.

Glossary

IQR	Interquartile range
AUC	Area under the curve
MAE	Mean absolute error
RMSE	Root mean squared error
AI	Artificial intelligence
CNN	Convolutional neural network
LR	Logistics regression
MLP	Multi-layer perceptron
FCL	Fully connected layer
PCA	Principal component analysis
EHR	Electronic health record
GUI	Graphical user interface
CT	Computed tomography
EEG	Electroencephalogram

Funding The study was supported by the Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna and the Italian neuroscience and neurorehabilitation research hospitals network ("Rete IRCCS delle Neuroscienze e della Neuroriabilitazione") which funded the study jointly with the "Ricerca corrente RC2020 program" and the 5 × 1000 funds AF2018: "Data Science in Rehabilitation Medicine" AF2019: "Study and development of biomedical data science and machine learning methods to support the appropriateness and the decision-making process in rehabilitation medicine" by the Italian Ministry of Health.

References

1. De Biase S, Cook L, Skelton DA, Witham M, ten Hove e R (2020) The COVID-19 rehabilitation pandemic. Age Ageing 49(5): 696–700. https://doi.org/10.1093/ageing/afaa118
2. Xie J et al (2020) Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med 46(5):837–840
3. Arabi YM, Murthy S, Webb S (2020) COVID-19: a novel coronavirus and a novel challenge for critical care. Intensive Care Med 46(5):833–836
4. LeBlanc M, Crowley J (1995) A review of tree-based prognostic models. Cancer Treat Res 75:113–124. ISSN: 09273042
5. Koutarou Matsumoto et al (2020) Stroke prognostic scores and data-driven prediction of clinical outcomes after acute ischemic stroke, pp. 1477–1483. Stroke. ISSN: 15244628
6. Leeuwenberg AM, Schuit E (2020) Prediction models for COVID-19 clinical decision making. The Lancet Digital Health 2(10):496–497. ISSN: 25897500
7. Siddique Latif et al (2020) Leveraging data science to combat COVID-19: a comprehensive review. IEEE Transactions on Artificial Intelligence, Early Access
8. Nguyen P, Tran T, Wickrmasinghe N, Venkatesh S (2016) Deepr: a convolutional net for medical records. arXiv
9. Yu C, Fei Wang F, Ping Zhang P, Jianying Hu J (2016) Risk prediction with electronic health records: a deep learning approach. SDM
10. Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng e M (2020) Reinforcement learning for clinical decision support in critical care: comprehensive review. J Med Internet Res 22(7):e18477, lug. https://doi.org/10.2196/18477
11. Liu Q, Fang X, Tokuno S, Chung U, Chen X, Dai X, Liu X, Xu F, Wang B, Peng P (2020) A web visualization tool using T cell subsets as the predictor to evaluate COVID-19 patient's severity. PLoS ONE 15(9):e0239695. https://doi.org/10.1371/journal.pone.0239695
12. Luke Moore Ahmed Abdulaal, Aatish Patel, Esmita Charani, Sarah Denny, Nabeela Mughal (2020) Prognostic modeling of COVID-19 using artificial intelligence in the United Kingdom: model development and validation. J Med Internet Research
13. Liu Y-P et al (2020) Combined use of the neutrophil-to-lymphocyte ratio and CRP to predict 7-day disease severity in 84 hospitalized patients with COVID-19 pneumonia: a retrospective cohort study. Ann Transl Med 8(10):635–635. https://doi.org/10.21037/atm-20-2372

14. Wang S et al (2020) A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 56(2):2000775, ago. https://doi.org/10.1183/13993003.00775-2020
15. Qi X, et al (2020) Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 Infection: a multicenter study», Infectious Diseases (except HIV/AIDS), preprint. https://doi.org/10.1101/2020.02.29.20029603
16. Syeda HB et al (2021) Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR Med Inform 9(1):e23811
17. World Health Organization (2020) Clinical management of severe acute respiratory infection when novel coronavirus (2019-nCov) infection is suspected: interim guidance. Available at: https://apps.who.int/iris/handle/10665/33089318
18. Arienti C, Campagnini S, Brambilla L, Fanciullacci C, Lazzarini S, Mannini A, Patrini M, Carrozza M (2021) The methodology of a "living" COVID-19 registry development in a clinical context. J Clin Epidemiol 142. https://doi.org/10.1016/j.jclinepi.2021.11.022
19. Breiman L (1998) Arcing classifier. Ann Stat 26(3):801–849
20. Breiman L (2001) Random forests. Mach Learn 45:5–32
21. Fukushima K (1998) Neocognitron: a hierarchical neural network capable of visual pattern recognition. Neural Netw 1(2):119–130
22. Khan A, Sohail A, Zahoora U, Qureshi e AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53(8):5455–5516. https://doi.org/10.1007/s10462-020-09825-6
23. Malek S, Melgani F, Bazi e (2018) One-dimensional convolutional neural networks for spectroscopic signal regression: feature extraction based on 1D-CNN is proposed and validated. J Chemom 32(5):e2977. https://doi.org/10.1002/cem.2977
24. Zhao J, Mao X, Chen e L (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Process Control 47:312–323. https://doi.org/10.1016/j.bspc.2018.08.035
25. Cruz RMO, Sabourin R, Cavalcanti e GDC (2014) On meta-learning for dynamic ensemble selection», in 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, ago. 1230–1235. https://doi.org/10.1109/ICPR.2014.221
26. Linn MW, Linn BS, Gurel L (1968) Cumulative Illness Rating Scale. J Am Geriatr Soc 622–626
27. Setti E, Liuzzi P, Campagnini S, Fanciullacci C, Arienti C, Patrini M, Mannini A, Carrozza MC (2021) Predicting post COVID-19 rehabilitation duration with linear kernel SVR. IEEE EMBS International Conference on Biomedical and Health Informatics. https://doi.org/10.1109/BHI50953.2021.9508602
28. Mannini A, Hakiki B, Liuzzi P, Campagnini S, Romoli A, Draghi F, Macchi C, Carrozza MC (2021) Data-driven prediction of decannulation probability and timing in patients with severe acquired brain injuries. Computer Methods and Programs in Biomedicine 209(4):106345. https://doi.org/10.1016/j.cmpb.2021.106345
29. Shamout F, Zhu T, Clifton D (2020) Machine learning for clinical outcome prediction. IEEE Rev Biomed Eng. 14:116–126
30. Nemati M, Ansary J, Nemati e N (2020) Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. Patterns 1(5):100074. https://doi.org/10.1016/j.patter.2020.100074
31. Ebinger J, Wells M, Ouyang D, Davis T, Kaufman N, Cheng S, Chugh S (2021) A machine learning algorithm predicts duration of hospitalization in COVID-19 patients. Intell Based Med 5:100035. https://doi.org/10.1016/j.ibmed.2021.100035
32. Chiari M, Gerevini AE, Maroldi R, Olivato M, Putelli L, Serina I (2021) Length of stay prediction for Northern Italy COVID-19 patients based on lab tests and X-ray data. Pattern Recognition. ICPR International Workshops and Challenges. https://doi.org/10.1007/978-3-030-68763-2_16
33. P.-F. (Jennifer) Tsai et al (2016) Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. J Healthc Eng 1–11. https://doi.org/10.1155/2016/7035463
34. Stone K, Zwiggelaar R, Jones P, Parthaláin e NM (2020) Predicting hospital length of stay for accident and emergency admissions. In Advances in computational intelligence systems, vol. 1043, Z. Ju, L. Yang, C. Yang, A. Gegov, e D. Zhou, A c. di Cham: Springer International Publishing, pp 283–295
35. PL is a PhD student at Fondaz. Don Gnocchi and Scuola Sant'Anna. With Biomed. Eng. bachelor and master in Neurom. Control, he works on cerebellar networks and ML for clinical outcome prediction
36. SC is a PhD student at Fondaz. Don Gnocchi and Scuola Sant'Anna. She had her bachelor in Mechatronics Eng. and master in Bionics Eng. Her interests are in ML and robotics in rehabilitation field
37. CF is a Psychologist, PhD, Clinical Trials Unit coordinator at IRCCS Fondaz. Don Gnocchi Florence, she monitors experimental protocols in central-southern area, according to ICH-GCP
38. CA is a PhD, Coordinator of Cochrane Rehabilitation and of the Clinical Trials Unit at IRCCS Fondazione Don Gnocchi-North Area. She works on CTs methodology in rehabilitation research
39. MP is a researcher at Fondazione Don Carlo Gnocchi and a General Practitioner. Since 2018, he has been part of Cochrane Rehabilitation Headquarters where he followed several projects
40. MCC is Prof. of Industrial Bioeng. at Scuola Sant'Anna coordinating the NeuroRobotics Area and President of CNR (Italian National Council of Research)
41. AM is Research Engineer at IRCCS Fondaz. Don Gnocchi and affiliate with Scuola Sant'Anna. His interests cover machine learning methods for signal processing and clinical outcome prediction