Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2023) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Recurrent Neural Networks for Daily Estimation of COVID-19 Prognosis with Uncertainty Handling

Nicholas Rossetti[a], Alfonso E. Gerevini[a], Matteo Olivato[a], Luca Putelli[a], Mattia Chiari[a], Ivan Serina[a], Davide Minisci[b], Emanuele Foca'[b]

[a] *University of Brescia, Via Branze, 38 Brescia, Italy*

[b] *Division of Infectious and Tropical Diseases, University of Brescia and ASST Spedali Civili General Hospital of Brescia,Piazzale Spedali Civili, 1 Brescia, Italy*

Abstract

Most ML-based applications for COVID-19 assess the general conditions of a patient trained and tested on cohorts of patients collected over a short period of time and are capable of providing an alarm a few days in advance, helping clinicians in emergency situations, monitor hospitalised patients and identify potentially critical situations at an early stage. However, the pandemic continues to evolve due to new variants, treatments, and vaccines; considering datasets over short periods could not capture this aspect. In addition, these applications often avoid dealing with the uncertainty associated with the prediction provided by machine learning models, potentially causing costly mistakes. In this work, we present a system based on Recurrent Neural Networks (RNN) for the daily estimate of the prognosis of COVID-19 patients that is built and tested using data collected over a long period of time. Our system achieves high predictive performance and uses an algorithm to effectively determine and discard those patients for whom RNN cannot predict the prognosis with sufficient confidence.

Keywords: COVID-19; Recurrent Neural Network; Gated Recurrent Unit; Time Distributed; Clinical Data; Uncertainty in ML

1. Introduction

The coronavirus pandemic caused by SARS-CoV-2 continues to be a global public health crisis. After more than three years of COVID-19 pandemic, more than 600 million confirmed infections and more than 6 million deaths related to COVID-19 have been reported around the world. During this period, COVID-19 spread throughout the

* Corresponding author: Nicholas Rossetti, nicholas.rossetti@unibs.it
E-mail address: {nicholas.rossetti, alfonso.gerevini, m.olivato, luca.putelli, m.chiari017, ivan.serina, emanuele.foca, d.minisci}@unibs.it

world in several waves with or without effective treatments or vaccines for different variants of the SARS-CoV-2 virus. Italy was the first western country afflicted by the COVID-19 pandemic and the Lombardy region suffered an increasing number of cases with a high rate of intrahospital mortality [22]. Brescia Hospital, a large hospital with 15709 beds, was one of the referral Hub Hospital according to the high number of hospitalised patients. In those emergency conditions, data scientists created many machine learning systems that deal with hospitalised patients with COVID-19 to predict the final outcome, the need for an intensive care unit, ventilation, etc. [3, 16, 23].

Apart from predictive performance, other important factors should be considered in order to evaluate the usefulness and applicability of such systems. First, most of them used data collected at the beginning of the pandemic that became less useful in the following years [1, 4, 10, 26] due to the *concept drift* [24]. This issue, caused by significant changes in the disease over time, undermines the effectiveness of many machine learning systems that cannot easily adapt to important changes in data distributions. Second, most approaches do not assess a patient day by day, identifying critical conditions as early as possible, but only provide a prediction at the end or a few days in advance of the patient's hospitalisation [13]. Finally, in very sensitive contexts, such as clinical practise, most applications are used without knowing when the predictions provided by their decision support system are not confident enough, leading to potentially costly mistakes.

In this work, considering the raw data of more than $6,000$ patients hospitalised at *ASST Spedali Civili of Brescia* collected from February 2020 to October 2021, we built an extended dataset containing demographic information and several standard laboratory test results that span across different phases of the COVID-19 pandemic: the initial emergency, the decrease in viral circulation, the spread of the Alpha and Delta variants, and the vaccination campaign. Then we propose a system based on Recurrent Neural Network (RNN) with Gated Recurrent Units that provides the outcome prediction (alive or decease) for each day of hospitalisation, facing many of the aspects introduced previously.

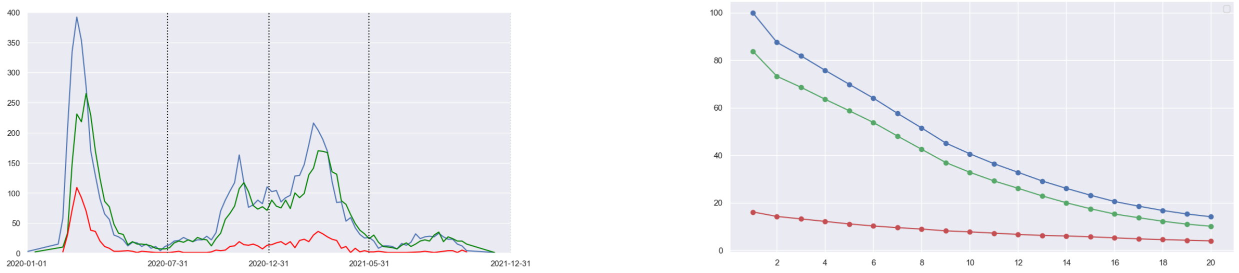
Regarding uncertainty, neural networks typically do not provide the confidence associated with a prediction, and standard techniques, such as calibration [15], are not particularly effective with healthcare datasets due to data quality issues [25]. Therefore, we have studied a method to introduce the uncertainty notion into our model. With the *target interpolation* technique, the RNN learns how to be more confident with the days passing, due to the availability of more data on the patients' conditions. Furthermore, due to the fact that our neural network does not have enough confidence to provide a definitive answer, we adapt our algorithm previously introduced in [10] to identify a threshold under which a patient is labelled as *unpredictable*, in order to avoid prediction errors, and to improve the overall performance of the system.

The rest of the paper is structured as follows. In Section 2 we discuss the related work. Then, in Section 3 we present the available data and our dataset, and in Section 4 we present our deep learning models. Finally, in Section 5 we present the results obtained by our models, in section 6 we discuss the ethics of the article and in Section 7 we report conclusions and future work.

2. Related work

In the scientific literature, there are many uses for RNNs to process time sequences, such as Electronic Health Records (EHRs) or clinical text [17, 19]. For instance, Lipton et al. [14] is one of the first studies that applies RNNs to predict patient diagnosis. The authors use the target replication technique: since the diagnosis is provided only at the end of the visits, they replicate this label to train the neural network for each timestamp. Choi et al. [6] developed Doctor AI, an RNN-based system that predicts the next visit's diagnosis and medication at each timestamp using a recurrent architecture applied to EHRs. However, in the context of the COVID-19 application, at the beginning of the pandemic, it was necessary to use machine learning techniques that required little training data, as it was very complex to obtain new data from hospitals given the emergency context. Yan et al. [13] use a dataset of blood samples from 485 affected patients in the Wuhan region, China, between January and February 2020 to train an XGBoost model to predict the prognostic risk in advance. Similarly, Booth et al. [1] predict the mortality risk of patients affected by COVID-19 using several laboratory chemistry parameters of 398 patients in early 2020 in Texas. Fernandes et al. [23] addressed several tasks, such as predicting admission to the ICU, ventilation, intubation, and death, for a dataset of more than 1000 patients from San Paolo, Brazil.

Taking into account studies that apply RNNs, Ramsy et al. [21] consider the clinical history of 200k patients before hospitalisation for COVID-19. Since the authors' goal is to estimate the prognosis and the need for a prolonged



(a) Number of hospitalised patients distributed over the years.

(b) Daily percentage of patients during hospitalisation.

Fig. 1: Patients hospitalised over the years (each vertical line represents the beginning of a new epidemic phase in Italy) and their daily distribution during the hospitalisation period. The blue line represents all patients, the red line represents dead patients, and the green line represents recovered patients.

hospital stay in the hospital triage phases, their study does not consider laboratory tests carried out during the hospitalisation period. In contrast, our study focusses on the evolution of these tests during hospitalisation. Fakhfakh et al. [8] apply a convolutional RNN to the sequence of X-ray images for the prediction of the prognosis of COVID-19 at the end of hospitalisation. Unfortunately, our dataset consists only of EHRs from laboratory tests without x-rays of the lungs. In Villegas et al. [26], the authors use a RNN to predict the patient's risk of death on a daily basis using the target replication technique. This Spanish study identifies the crucial days for the prognosis by exploiting attention's weights using the label replication technique. Their datasets contain data from the start of the pandemic through the 2020 fall. However, they do not provide any method for dealing with the uncertainty related to the prediction, and they consider a shortened data period limited to the first pandemic phase.

Various methods in the literature have been proposed to evaluate uncertainty estimation. Monte Carlo Dropout [9] is a technique that interprets the use of the dropout as a Bayesian approximation of a probabilistic model. Calibration [15] is a technique that uses sigmoid or isotonic regression to calibrate the probabilities of an already trained classifier. Deep Ensembles [12] leverage the ensemble concept to create different neural networks to get a better estimate of their confidence. Finally, NGBoost [7] exploites the natural gradient to perform gradient boosting by casting it as a problem of determining the parameters of a probability distribution over a base learner.

Using only laboratory tests for a dataset of approximately 2000 patients, we conducted a preliminary study using ensembles of decision trees, at different times during patient hospitalisation, to estimate the prognosis [10]. One of the key aspects of [10] is that we introduced an algorithm to calculate a threshold under which uncertain predictions could be excluded. Furthermore, considering the same data, in [4] we trained a Many-to-one RNN using the sequence of clinical laboratory exams to predict the patients' prognosis of COVID-19. We implemented data pre-processing techniques (such as cutting strategy and dataset augmentation) to predict prognosis in advance.

In the present work, we extend and improve our previous analyses in several key aspects, such as providing a prediction for every day of hospitalisation with a Many-to-Many RNN and introducing a new approach to handle uncertainty related to model's decisions. Furthermore, we consider a significantly extended dataset [16] that spans across several waves of the pandemic using data collected over two years, also addressing possible concepts drift issues [24].

3. Available Data and Dataset Creation

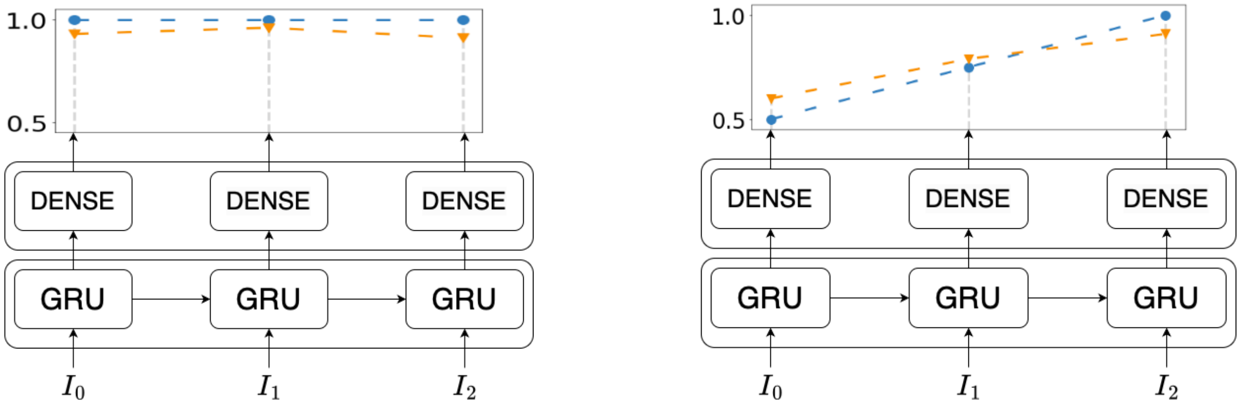
In this work, we consider 6435 patients hospitalised at *ASST Spedali Civili of Brescia* for COVID-19 from March 2020 to December 2021. During this long period of time, we witnessed an evolution of the disease in its variants during the two years of the pandemic: clinicians gained more knowledge and experience, new treatments were discovered; new variants of the Sars-Cov-2 virus emerged and spread throughout the world, and last but not least, vaccines were developed and inoculated to the vast majority of people. Therefore, we analyse the data available by dividing them into different periods and waves. In Figure 1a, we report the trend of hospital admissions (*blue line*), discharge from hospitalisation (*green line*) and deaths of patients (*red line*) affected by COVID-19 in the hospital. Patients are divided into four different groups (divided by a vertical line in Figure 1a), according to the date of hospitalisation. This is due to the different pandemic waves that the hospital and Italy were facing[1]:

- **First Wave** (January 2020 to July 2020): from March 2020, the virus rapidly spread in Italy, causing hospitals overload. For example, our structure hospitalised more than 250 COVID-19 patients a day for more than a month. Due to the limited healthcare resources there was no formal distinction between the admission wards as the hospital tried to maximise the number of admitted patients. From June 2020, the virus circulation decreased, with almost no deaths and very few hospitalisations.
- **Second Wave** (August 2020 to December 2020): the virus circulation increases during the autumn period, caused by the resumption of work and school after summer break. Therefore, another wave hits our region, causing a hospital emergency and new restrictions were introduced. However, despite the increase of the hospitalisations, the number of deceased patients was lower relative to the first wave.
- **Third Wave** (January 2021 to May 2021): at the beginning of 2021, hospitalisation and mortality increased due to diffusion of the SARS-CoV-2 Alpha variant. However, from April 2021, hospitalised patients also started to decrease rapidly due to an effective vaccination campaign.
- **Fourth Wave** (June 2021 to December 2021): despite the high number of cases in Italy, hospital admissions and mortality remained low thanks to a useful vaccination campaign. In this period, the Delta Variant became dominant.

During hospitalisation, clinicians perform various biochemical and virological tests to assess patients' conditions (improving or worsening) and evaluate the transfer to intensive care. The features considered can be divided into static, i.e. fixed for the entire sequence, and dynamic, i.e. variable during the time series. The static features considered are patients' sex and age. Dynamic features are the exam results (PCR, LDH, Ferritin, Troponin-T, WBC, D-Dimer, Fibrinogen, Lymphocytes and Neutrophils, the COVID-19 rinopharingeal swab), the hospital ward in which the patient stays, and a boolean flag representing if a patient has been admitted to the intensive care unit. We do not have additional information on symptoms, comorbidities, generic health conditions, or clinical treatments. Another problem of an emergency period is the irregular frequency of examinations during hospital admission due to different health conditions and the availability of various machines. As seen in Figure 1a, the COVID-19 virus in these two years has outlined various pandemic phases caused by the different variants, circulation, new treatment methods, and vaccination campaign. Following the methodology introduced in [10], we use the entire dataset with an additional feature that helps the learning algorithm discriminate whether hospitalisation is or not during a particularly critical phase. The new feature, called Death Rate, aims to provide an indicator of the state of the pandemic emergency on a given day and is defined as the average death rate calculated considering the seven days preceding that day. Specifically, the death rate feature is the ratio of all patients who died over all patients discharged (dead or alive) during the seven days considered.

In the following, we describe how we build the datasets from training and testing our machine learning models, starting from the raw data related to the considered cohort of patients. We designed a representation of the patient's stay over time. As introduced in [4], we create a matrix $M[l, e]$, where l is the number of days of hospitalisation and e is the number of features and lab tests considered. We impute the features day by day (the value obtained for the test j on the i^{th} day of hospitalisation is stored in $M[i, j]$) and replicate static ones. If a value is missing for a given day (i.e. the exam has not been performed), we insert -3 in place of the missing value (i.e. $M[i, j] = -3$ with the test j on day i). More complex imputation methods, such as mean, median, MICE, Miss-Forest and Fill-Forward, were tested and rejected as they did not lead to an improvement in the performance of the models. After building the matrix, we scale and normalise our data using the Standard Scaler [20] and divide them by stratified sampling into 80% patients to train the models and 20% to test partitioning using label, age and gender. Figure 1b shows the number of patients recovered for each day. From the dialogue with the medical staff, the average hospitalisation of a patient for COVID-19 is between one and two weeks. For this reason, we have selected the hyperparameter $l = 20$. The figure shows this behaviour within our dataset, after two weeks, about 80% of the patients have been discharged from the hospital. For those patients hospitalised for more than this length, in agreement with the medical staff, the

[1] National COVID-19 data available at: https://github.com/pcm-dpc/COVID-19



(a) Target Replication Network

(b) Target Interpolation Network

Fig. 2: Models of the network of a deceased patient hospitalised for 3 days. The blue line represents the patient's label, the orange line the prediction.

last 20 days were considered in order to evaluate the laboratory tests closest to the end of the hospitalisation. From the total number of patients (6435) there are 5394 recovered patients and 1041 dead patients. During hospitalisation, the daily ratio between healed and deceased patients settles at an average value of 15.47% and a standard deviation of 5.

4. Day-by-day Prediction Models

In this section, we describe our models, the **Target Replication Network** and the **Target Interpolation Network**, to estimate the prognosis of patients (released alive or deceased) every day during their hospitalisation.

Both models are based on RNNs and monitor the mortality risk from hospital admission to discharge. Our aim is to provide a day-by-day prediction of the hospitalisation outcome with the goal of anticipating it as soon as possible in order to allow medical personnel to intervene promptly. As shown in Fig. 2, the recurrent layer consists of a unidirectional Gated Recurrent Unit (GRU) layer [5], producing an output value for each hospitalisation day. The GRU processes the new input data (i.e., the lab tests conducted in each day) alongside the most relevant information from the previous days. The GRU output of each day is, in fact, processed by a Time-Distributed feed-forward layer, which ultimately provides the prognosis estimation. Please note that we do not use a different feed-forward layer for each day, but the weight matrix is shared across the whole data sequence.

Despite these similarities, the Target Replication and Target Interpolation networks present several important differences, which we explain in the following sections.

Our dataset is unbalanced (5394 recovered and 1041 dead patients), and our main interest is to identity the most critical patients (these need prompt treatment to prevent death). For this reason, we cannot evaluate our model using accuracy, because we can only predict the majority class and obtain very high values through this score. We measure the performance of our models using F_2 and ROC-AUC metrics. ROC-AUC is a widely used metric in the medical field and is easy to understand for medical personnel.

4.1. Target Replication Network (TRNet)

TRNet considers the estimation of the prognosis as a binary classification task. In this configuration, the label is set to 0 if the patient is going to be released alive, 1 in case of decease. Although this is known only at the end of the patient's stay, with the target replication technique [14], the label is replicated on each hospitalisation day at the training time. For instance, if a patient is going to be released alive after 12 days of stay, the RNN and the Time-Distributed Layer have the goal of predicting the label 0 twelve times, one prediction for each day. Using TRNet, the user can obtain a prediction without delay and allow clinicians to intervene promptly. A schematic representation of this model can be seen in Fig. 2a considering a deceased patient.

In TRNet, the Time-Distributed layer is made by one neuron with the sigmoid activation function. Therefore, for a patient who stays n days, the output of the network is a vector of n elements between 0 and 1. For each given day, we consider the patient at risk of death if the output value is greater than 0.5. To train the network, we use F-β loss [11] which is often used in unbalanced contexts and allows, through the choice of β, to penalise more the errors on the minority class if the network makes a mistake in its training. F_2-Loss is the best loss obtained in the search for hyperparameters, other losses for unbalanced contexts have been discarded due to worse performance (weighted cross entropy, F_1-Loss and Focal Loss). We set $\beta = 2$ to weight more the prediction errors made for patients at risk of death, which is the minority class within our dataset (16% of the total). In fact, not identifying a serious patient at an early stage is very costly and this mistake could cause death. The formula is: $F_\beta\text{-Loss} = 1 - \frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2 * Precision + Recall + \varepsilon}$, with $\beta = 2$, and $\varepsilon = 10^{-7}$ used to avoid zero-division errors.

4.2. Target Interpolation Network (TINet)

To avoid costly mistakes, we must consider the uncertainty related to the predictions made by the RNN-based model. Given that we predict a value between 0 and 1, we could assume that when the network predicts a value about 0.5, it is fundamentally uncertain and the final decision is almost given randomly. Instead, for higher values, the network should have higher confidence. However, we have experimentally verified that the TRNet model, due to its training configuration, often produces outputs very close to 0 (i.e., high confidence that the patient will be released alive) or very close to 1 (very probable decease), even on the first days of hospitalisation and also in case of errors. Therefore, in TINet we tried to model and introduce this uncertainty with an innovative technique that we call **Target Interpolation**. The insight behind this approach is that the patient's condition is uncertain at triage time and becomes clearer as the release date approaches. The target interpolation is implemented as follows. On the first day, we consider that there is the maximum uncertainty about the patient's conditions. Therefore, we set the label to 0.5. Instead, at the end of hospitalisation, we consider that we are fully certain about the final outcome of the patient; therefore, we set the label at 0 if he is discharged or 1 in the case of decease. The labels for the intermediate days are calculated by a linear interpolation. For instance, for a patient who is going to be discharged in 6 days, the 6 labels are going to be 0.5, 0.4, 0.3, 0.2, 0.1 and 0. On the contrary, for deceased patients, the labels will gradually increase to 1. With this change, the problem changes from a classification task to a regression task. Therefore, while the Time Distributed layer has the same structure as the TRNet, for training the TINet we use the Mean Squared Error Loss function. In order to enhance the minority class in the search for hyperparameters we use F_2 metric as a configuration evaluation function.

With this change, our neural network can understand that more time passed in the hospital and more lab tests are performed lead to a better understanding of the patient's conditions in accordance with clinicians, and therefore a greater certainty in prognosis estimation. Moreover, we can select a threshold under which we can discard uncertain predictions. In this way, the patient can also be labelled as *unpredictable*, alerting the clinicians and possibly avoiding costly mistakes.

To find this threshold, we take advantage of an adapted version of the FindUncertainThreshold algorithm introduced in [10]. At training time, the algorithm iteratively selects a threshold (starting with the lowest probability of prediction) and removes uncertain patients. If the number of patients removed exceeds a predefined maximum percentage of discarded patients, then the algorithm terminates by returning the last valid threshold. Otherwise, it calculates the performance with the remaining group of patients. If the results obtained are better than the previous ones, we save the new value and continue until we reach the maximum predicted probability. We performed the whole procedure in ten-fold cross-validation to learn a threshold using only training data. The threshold is then applied to the test set. However, w.r.t. the first version described in [10], we have a significant difference. Although the original algorithm was designed for a single output, this new version provided a threshold for each given day. Note that the maximum percentage of patients labelled uncertain is an hyperparameter. According to the clinicians, we select 25%, but other thresholds can be considered. Other standard techniques for evaluating uncertainty, such as calibration [15] or Monte-Carlo Dropout [9], did not produce acceptable results. NGBoost [7] and Deep Ensemble [12] were not applicable in our context.

Model	Days of hospitalisation											
	2		4		8		12		16		20	
	RA	F_2	RA	F_2	RA	F_2	RA	F_2	RA	F_2	RA	F_2
Snapshot [10]	0.80	0.69	0.81	0.70	0.79	0.67	0.81	0.72	0.80	0.72	0.80	0.76
Many-to-One [4]	0.75	0.66	0.78	0.70	0.76	0.65	0.84	0.73	**0.92**	**0.87**	**0.97**	**0.95**
TINet-NoThreshold	0.80	0.68	0.80	0.67	0.85	0.76	0.86	0.78	0.89	0.83	0.94	0.91
TRNet-NoThreshold	**0.81**	**0.70**	**0.84**	**0.74**	**0.86**	**0.77**	**0.87**	**0.80**	0.90	0.85	0.95	0.92
Snapshot-Threshold [10]	**0.89**	**0.81**	**0.87**	**0.77**	0.88	0.80	0.87	0.81	0.85	0.80	0.88	0.86
TINet-Threshold	0.85	0.75	0.85	0.74	**0.91**	**0.85**	**0.92**	**0.86**	**0.94**	**0.90**	**0.97**	**0.95**

Table 1: Performance of our models in terms of ROC-AUC (RA) and F_2 score on different days during the patient's hospitalisation. In the first part of the table (above the horizontal line) the models without the threshold. Below the line, models with the threshold.

5. Experimental Results

In this section, we report the results obtained by applying Target Replication and Target Interpolation techniques to time-distributed RNNs. For the latter, we also report the results of the model, including also the FindUncertainThreshold algorithm to remove uncertain patients. TRNet and TINet models are trained using Adam as the optimiser and performing a hyperparameter search using the Bayesian approach provided by the Optuna framework [2]. Hyperparameters (such as imputers, scalers, number of hidden levels, cell typology between LSTM and GRU, the number of neurons, dropout, recurrent dropout, batch size and losses) are evaluated in 10-fold cross-validation. To avoid false negatives (i.e., patients whose risk of decease is not identified), the configuration that obtains the best F_2 score is selected. To avoid overfitting, we implement recurrent dropout, batch normalisation, activity regulariser, and early stopping techniques.

In order to show the efficacy of our techniques, we made a comparison with three different baselines on different days during the patient's stay in terms of ROC-AUC and F_2:

- the **Snapshot** approach introduced in [10, 16], into which an ensemble of decision trees (Random Forest, Extra-Trees, XGBoost or LightGBM) is trained considering the demographic information, the last available lab tests and some additional trend features for describing the evolution of the patient's conditions; in this configuration, one model is trained for each day.
- the **Thresholded Snapshot**, considering the same model as the previous approach but applying the FindUncertainThreshold [10] algorithm to each model;
- the **Many-to-one** approach based on RNNs, introduced in [4], in which the label is only at the end of hospitalisation.

Since public datasets on hospitalised patients for COVID-19 are not present in the literature, it was not possible to use other baselines in addition to the old models already published on this dataset. Please note that although the test set is made up of the same cohort of patients, in evaluating the n^{th} day, patients who stayed in the hospital for less than n days are excluded from the test set.

In Table 1, we report our results in terms of ROC-AUC and F_2 of our networks. In terms of ROC-AUC TINet (with Threshold) reaches a value higher than 0.9 from day 8. With respect to TRNet, the improvement is by several points, especially from day 10 to day 18 while F_2 metric behaves similar to ROC-AUC with slightly lower values (reaches a value higher than 0.9 from day 16). This is mainly due to the use of FindUncertainThreshold, which improves performance by labelling some patients as *unpredictable*, reducing costly mistakes. This can be seen also by analysing the TINet (No Threshold) model, which is significantly lower w.r.t. both TRNet and TINet, due to the negative impact on introducing the uncertainty in the model training. An important aspect is that as the number of days of hospitalisation increases, both the confidence in network prediction and the threshold value increase. In fact, the threshold value is around 0.55 at the beginning of hospitalisation and reaches around 0.7 at the end. Furthermore, the actual number of patients labelled unpredictable is generally between 15% and 20% and always less than 25%, as

Model	Days of hospitalisation									
	2	4	6	8	10	12	14	16	18	20
TINet first wave	**0.84**	0.83	**0.86**	0.85	**0.93**	0.91	0.90	0.89	0.94	0.92
TINet second wave	0.80	0.84	0.85	0.87	0.85	0.88	0.91	0.94	**1.00**	0.98
TINet third wave	0.80	0.84	**0.86**	**0.93**	**0.93**	**0.94**	**0.96**	0.96	0.97	0.95
TINet fourth wave	0.78	**0.89**	**0.86**	0.85	0.84	0.86	0.86	**1.00**	**1.00**	**1.00**

Table 2: Performance of TINet model on the various waves in terms of ROC-AUC score over the different epidemic waves.

explained in Section 4.2. Please note that applying FindUncertainThreshold to TRNet is an incorrect operation. In fact, given the training configuration into which the label is always 0 or 1, the network tends to predict values very close to 0 or 1 which cannot be interpreted as a highly confident prediction but only as a label approximation. From a clinical point of view, although the two methods applied have shown similar performances, TINet is more reliable since in clinical practise it is difficult to apply a binary model when the patients are preestablished to be alive or dead.

In Table 1 we also show the comparison between the networks and the baselines in terms of ROC-AUC and F_2, day by day. Compared to baselines, our models perform better than snapshot approaches (thresholded or not) and Many-to-one RNN, especially after 6 days of hospitalisation or more. Given the threshold-free models, TRNet is the best model, already from the first days the ROC-AUC and F_2 values surpassing by a few points the ensembles of trees and the Many-to-One network. The only exception is at the end of hospitalisation, where the many-to-one network catches up with and narrowly exceeds TRNet since it has been trained only to predict the patient's diagnosis at the end of the hospitalisation. Analysing also the models with threshold, the best model is TINet which obtains the best results in terms of ROC-AUC and F_2 for the entire hospitalisation period. However, there are two notable exceptions: the Thresholded Snapshot models perform slightly better than TINet for days 2 and 4. This is probably due to the fact that snapshot models are trained specifically for that day, while the RNN (which is trained only once) could encounter some difficulties in dealing with very small sequences that often contain missing values. This can also be seen for the Many-to-One RNN, which reaches acceptable performance (and is always lower w.r.t. TINet) only after 10 days of hospitalisation or more. For prolonged stays, our approaches can exploit the information contained in the entire patient's clinical history, achieving very high performance. In fact, TINet almost reaches a score of 1 in terms of ROC-AUC.

In general, our RNN models provide better or comparable results w.r.t. the ensemble of decision trees while maintaining greater flexibility. In fact, tree-based models are trained on a specific day and require a model for each analysis day. Thus, these models maintain a constant performance throughout the sequence, without improvement. Instead, we can train a single RNN for the entire hospitalisation period of the patient and provide a daily prediction, obtaining an improvement as hospitalisation of the patient increases.

To analyse the performance of our models on the progression of the epidemic, we divided our test set into four parts, one for each stage we described in Section 3 and evaluated the performance of TINet (with Threshold) in terms of ROC-AUC for these four groups of patients. Note that the model is trained on data from across the pandemic and only test data was split in order to perform this analysis. The results are shown in Table 1, we can observe how the model obtains comparable results in the various phases of the pandemic, providing more than acceptable results with a ROC-AUC often higher than 0.8 despite the different variants, treatments, and vaccines. However, there is some variability (6 points of standard deviation). The best results are obtained in the first and third phases, reaching values above 0.9 on the 10^{th} day, where the number of hospitalised patients is higher. In these phases, this behaviour is related to the use of F_2, which is more focused on critically ill patients in the model. On the other hand, the model has a slower ROC-AUC growth that exceeds 0.9 in the other two waves only after the 15^{th} day. These small issues in the milder phases could be due to the low number of patients admitted in those periods; therefore, limiting the learning capabilities of the models for those periods. We have manually observed that in these cases, the model tends to predict more critically ill patients, as if it is in an emergency, worsening performance. Generally, the number of false negatives remains very low, avoiding very dangerous mistakes.

6. Ethics

The present observational retrospective study was approved by Ethic Committee of Brescia Province, Italy on 17th January 2022 (number 5149/2022), in accordance with current regulations (Legislative Decree no. 211 of June 24, 2003 and subsequent additions and authorisations), carried out in full respect of human dignity and fundamental rights as dictated by **Declaration of Helsinki**, by the standards of **Good Clinical Practise**. Patient data have been made anonymous (alpha-numeric code) in observance of the rights provided for by privacy legislation (Legislative Decree no. 196/2003 Art. 7).

7. Conclusions and Future Work

We have presented a system to assess the mortality risk of hospitalised patients with COVID-19 daily, considering the results of laboratory tests from more than 6000 patients during almost two years of the pandemic. We built two RNN-based models that exploit different label training mechanisms to predict prognosis day by day: the Target Replication Network and the Target Interpolation Network. The last model, applying the FindUncertainThreshold algorithm, can identify patients for which the model is not confident enough to provide a reliable prediction. Our experimental results show that TRNet obtains a high ROC-AUC score from the start of the hospitalisation to its end. Applying the FindUncertainThreshold algorithm to TINet, we show better performance through hospitalisation, excluding uncertain patients. We divided our test set into various pandemic phases and analysed the behaviour of the TINet (with threshold) during the evolution of the virus. We show that the model works very well in epidemiological peaks, while it tends to present a more serious situation than reality in mild intervals.

In clinical practise, these experiments have different applicability. First, it is important to support physicians with a new and useful tool to predict the clinical evolution of a severe disease. Furthermore, this tool is very important to assess the probability that a patient will worsen or die by taking appropriate diagnostic and therapeutic precautions. However, from a health policy point of view, these models can be helpful in estimating the degree of saturation of hospital wards (both medical and intensive care units), the severity of admitted patients, and therefore estimating the patient's prognosis and the probable duration of hospitalisation.

As a future work, we plan to analyse the explainability of our models to identify the most important features and extract a simple explanation of the main criteria used by them and to test the use of Transformer-based models [18]. Moreover, these results could be applied not only to COVID-19 but also to other infectious diseases both acute and chronic where the prediction and estimation of the probability of worsening and death are crucial for the clinical management of the patients.[2]

References

[1] Adam, B., Elizabeth, A., Peter, M.: Development of a prognostic model for mortality in covid-19 infection using machine learning. Modern Patology (2021)

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., Karypis, G. (eds.) Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019. pp. 2623–2631. ACM (2019). https://doi.org/10.1145/3292500.3330701

[3] Chiari, M., Gerevini, A.E., Maroldi, R., Olivato, M., Putelli, L., Serina, I.: Length of stay prediction for northern italy COVID-19 patients based on lab tests and x-ray data. In: Pattern Recognition. ICPR International Workshops and Challenges - Proceedings, Part I. Lecture Notes in Computer Science, vol. 12661 (2020)

[4] Chiari, M., Gerevini, A.E., Olivato, M., Putelli, L., Rossetti, N., Serina, I.: An application of recurrent neural networks for estimating the prognosis of COVID-19 patients in northern italy. In: Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021. Lecture Notes in Computer Science, vol. 12721 (2021)

[5] Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, (2014)

[2] This work was carried out while Nicholas Rossetti was enrolled in the Italian National Doctorate on Artificial Intelligence, conducted by Sapienza University of Rome in collaboration with University of Brescia.

[6] Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor AI: predicting clinical events via recurrent neural networks. In: Proceedings of the 1st Machine Learning in Health Care, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016. JMLR Workshop and Conference Proceedings, vol. 56, pp. 301–318. JMLR.org (2016)

[7] Duan, T., Avati, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A.Y., Schuler, A.: Ngboost: Natural gradient boosting for probabilistic prediction. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 2690–2700. PMLR (2020)

[8] Fakhfakh, M., Bouaziz, B., Gargouri, F., Chaâri, L.: ProgNet: COVID-19 Prognosis Using Recurrent and Convolutional Neural Networks. The Open Medical Imaging Journal **12**(1), 11–12 (Dec 2020)

[9] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 1050–1059. JMLR.org (2016)

[10] Gerevini, A.E., Maroldi, R., Olivato, M., Putelli, L., Serina, I.: Prognosis prediction in covid-19 patients from lab tests and x-ray data through randomized decision trees. In: Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020. CEUR Workshop Proceedings, vol. 2675 (2020)

[11] Kawahara, J., Hamarneh, G.: Fully convolutional neural networks to detect clinical dermoscopic features. IEEE J. Biomed. Health Informatics **23**(2) (2019)

[12] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 6402–6413 (2017)

[13] Li, Y., Hai-Tao, Z., Jorge, G., Yang, X., Maolin, W., Yuqi, G., Chuan, S., Xiuchuan, T., Liang, J., Mingyang, Z., Xiang, H., Ying, X., Haosen, C., Yanyan, C., Tongxin, R., Fang, W., Yaru, X., Sufang, H., Xi, T., Niannian, H., Bo, J., Cheng, C., Yong, Z., Ailin, L., Laurent, M., Junyang, J., Zhiguo, C., Shusheng, L., Hui, X., Ye, Y.: An interpretable mortality prediction model for covid-19 patients. Nature Machine Intelligence (2020)

[14] Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.C.: Learning to diagnose with LSTM recurrent neural networks. In: 4th International Conference on Learning Representations, ICLR, Conference Track Proceedings (2016)

[15] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Raedt, L.D., Wrobel, S. (eds.) Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005. ACM International Conference Proceeding Series, vol. 119, pp. 625–632. ACM (2005)

[16] Olivato, M., Rossetti, N., Gerevini, A.E., Chiari, M., Putelli, L., Serina, I.: Machine learning models for predicting short-long length of stay of covid-19 patients. Procedia Computer Science **207** (2022), proceedings of the 26th International Conference KES2022

[17] Putelli, L., Gerevini, A.E., Lavelli, A., Maroldi, R., Serina, I.: Attention-based explanation in a deep learning model for classifying radiology reports. In: Tucker, A., Abreu, P.H., Cardoso, J.S., Rodrigues, P.P., Riaño, D. (eds.) Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12721, pp. 367–372. Springer (2021)

[18] Putelli, L., Gerevini, A.E., Lavelli, A., Mehmood, T., Serina, I.: On the behaviour of bert's attention for the classification of medical reports. In: Musto, C., Guidotti, R., Monreale, A., Semeraro, G. (eds.) Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence(AIxIA 2022), Udine, Italy, November 28 - December 3, 2022. CEUR Workshop Proceedings, vol. 3277, pp. 16–30. CEUR-WS.org (2022)

[19] Putelli, L., Gerevini, A.E., Lavelli, A., Olivato, M., Serina, I.: 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. Procedia Computer Science, vol. 176, pp. 349–359. Elsevier (2020)

[20] Raju, V.N.G., Lakshmi, K.P., Jain, V.M., Kalidindi, A., Padma, V.: Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (2020)

[21] Rasmy, L., Nigo, M., Kannadath, B.S., Xie, Z., Mao, B., Patel, K., Zhou, Y., Zhang, W., Ross, A., Xu, H., Zhi, D.: Recurrent neural network models (covrnn) for predicting outcomes of patients with covid-19 on admission to hospital: model development and validation using electronic health record data. The Lancet Digital Health **4** (2022)

[22] Rizzi, M., Castelli, F., Latronico, N., Focá, E.: Sars-cov-2 invades the west. how to face a covid-19 epidemic in lombardy, northern italy? Infez Med (2020)

[23] Timoteo, F.F., de Oliveira Tiago Almeida, Esteves, T.C., de Moraes, B.A.F., Gabriel, D.C., Porto, C.F.A.D.: A multipurpose machine learning approach to predict covid-19 negative prognosis in são paulo, brazil. Scientific Reports (2021)

[24] Uchida, T., Yoshida, K.: Concept drift in japanese covid-19 infection data. Procedia Computer Science **207** (2022), proceedings of the 26th International Conference KES2022

[25] Van Calster, B., McLernon, D., van Smeden, M., Wynants, L., Steyerberg, E.: Calibration: The achilles heel of predictive analytics. BMC Medicine **17** (2019)

[26] Villegas, M., Gonzalez-Agirre, A., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Carrino, C.P., Pérez-Fernández, D., Soares, F., Serrano, P., Pedrera, M., García, N., Valencia, A.: Predicting the evolution of covid-19 mortality risk: A recurrent neural network approach. Computer Methods and Programs in Biomedicine Update **3** (2023)

Commentary

We would like to thank both reviewers for their insightful comments on the paper, as these comments led us to an improvement of the work. With respect to the first version of the paper, we have taken the reviews into account and made the following changes:

- we have improved the presentation and the writing as requested by the reviewers;
- we have improved the explanation of how training, validation and tests were selected;
- we have modified the keywords, including Recurrent Neural Networks as full text.