Predicting respiratory failure in patients with COVID-19 pneumonia: a case study from Northern Italy

Davide Ferrari[1] and **Federica Mandreoli**[2] and **Giovanni Guaraldi**[3] and **Jovana Milić** [4] and **Paolo Missier**[5]

Abstract.

The Covid-19 crisis caught health care services around the world by surprise, putting unprecedented pressure on Intensive Care Units (ICU). To help clinical staff to manage the limited ICU capacity, we have developed a Machine Learning model to estimate the probability that a patient admitted to hospital with COVID-19 symptoms would develop severe respiratory failure and require Intensive Care within 48 hours of admission. The model was trained on an initial cohort of 198 patients admitted to the Infectious Disease ward of Modena University Hospital, in Italy, at the peak of the epidemic, and subsequently refined as more patients were admitted. Using the LightGBM Decision Tree ensemble approach, we were able to achieve good accuracy (AUC = 0.84) despite a high rate of missing values. Furthermore, we have been able to provide clinicians with explanations in the form of personalised ranked lists of features for each prediction, using only 20 out of more than 90 variables, using Shapley values to describe the importance of each feature.

1 Introduction

In this paper we report on a machine learning exercise using an evolving, unstable, and limited training set, aimed at supporting hospital clinical staff during the COVID-19 crisis in Italy. The pandemic evolved very rapidly over the course of a few weeks, with Italy recording the first severe clusters of virus spread in Europe between February and March, 2020. This put frontline health services into emergency mode, forcing them to adapt very rapidly to an overload of patients with severe complications, primarily viral pneumonia. In addition to the clinical challenges, hospital medical staff had to deal with a shortage of critical care resources, mainly Intensive Care Unit (ICU) beds.

At the University Hospital in Modena, Italy (UHM), this translated into the urgent need to rapidly design end deploy new protocols for recording medical records, which had to integrate routine patient assessment information, eg blood tests, with observations about their complications (respiratory issues), a record of patients transfers across departments, namely the Infectious Diseases clinic, the ICU, and a number of clinics to deal with specific complications. Doctors also started recording details of tentative therapies, whose effectiveness was largely unknown at the time.

A new patients' database was created, however this was subject to a continuously evolving schema, with the result that the hundreds of precious daily data points exhibited sparsity problems, i.e., when a new variable was introduced and not retrospectively populated for the existing patients, as well as heterogeneity and inconsistencies.

Against this backdrop, there was an immediate need for analytics that would help clinicians to answer some of the more pressing questions. These included, besides the more clinical questions on the efficacy of therapies, the challenge of predicting the needs for limited ICU resources within a limited time horizon.

We focused specifically on the problem of predicting whether a patient would develop acute respiratory distress syndrome (ARDS) [13], leading to moderate to severe respiratory failure within hours, and thus to the need for assisted breathing and to admin the patient to ICU. This question translates well into a quantitative outcome that can be used in machine learning, namely by measuring the respiratory rate (a critical cutoff is > 30 breaths per minute), blood oxygen saturation $< 93\%$, and more importantly, the PaO_2/FiO_2 ratio of partial pressure of arterial oxygen to the fraction of inspired oxygen. The reach of a clinically-defined cutoff (150 mmHg) in at least one of two two consecutive days after 48 after admission, was taken as the proxy measure of choice to put the patient in a critical state where the need for assisted breathing was assessed. Clearly, the ability to predict this outcome with some advance notice would give medical staff information for manging ICU resources. From a machine learning persective, this is a well-defined binary classification problem, where the respiratory condition becomes a binary outcome that can be evaluated on the ground training set represented by the patients' database.

In the rest of the paper we outline the the challenges associated with the machine learning task, we describe the approach and evaluate the results, and outline the direction of our ongoing research.

1.1 Challenges and requirements

The specific context around data collection and curation, as well as the urgent need to manage ICU resource allocation, translated into a number of requirements and technical challenges.

Firstly, the dataset has been evolving rapidly not only in number of records but also, critically, in the schema, with new attributes added most daily as the requirements of downstream analysis became increasingly clear. As explained in more detail in Sec. 2, the dataset consist of Electronic Health Records (EHR) including routinely collected clninical tests, but also ad hoc observations, associated with the specific symptoms. Existing EHR collection systems were there-

[1] Università di Modena e Reggio Emilia, IT, email: davideferrari@unimore.it
[2] Università di Modena e Reggio Emilia, IT, email: federica.mandreoli@unimore.it
[3] Università di Modena e Reggio Emilia, IT, email: giovanni.guaraldi@unimore.it
[4] Università di Modena e Reggio Emilia, IT, email: jovana.milic@unimore.it
[5] Newcastle University, UK, email: paolo.missier@newcastle.ac.uk

fore inadequate. A consequence of this predicament is a continuously changing set of features, some of which require experts' explanations, which complicates the learning process.

Secondly, different labs may operate different practices and adopt slightly different standards, including different laboratory biomarkers. Care must therefore be taken when aggregating their values, as in general their source would have to be taken into account. Furthermore, different types of data are collected with varying frequency over time. Some systematically, some only on demand, and some, such as information about co-morbidities are collected only once, but not for every patient.

Related to this is the problem of data sparsity and of missing data. One example is the *Interleukin 6* variable, which was found to be relevant for analysis only after weeks of lockdown, thus is absent for a substantial proportion of the patients. While missing data can sometimes be inferred, or imputed, from available value distributions, this was not an option when dealing with critical patients vital parameters, which by their own nature are subject to abrupt changes and thus should not be exrapolated from known distributions. In fact, one may argue that the value of the data in this context is the change in data values, signalling an impending crisis.

Thus, a key challenge for the model is to be robust to missing data, a property that is not enjoyed by the majority of the available off-the-shelf libraries.

We should also mention that two milestone data extraction processes took place, with intermediate data versions in between, as explained in the next Section. The characteristics of the dataset changed with respect, for instance, to class imbalance and data sparsity, requiring different modelling strategies.

A separate challenge concerns the **trustworthiness** and **transparency** of the model itself, both of which are required if the model is to be embraced in clinical practice. As a general principle in Machine Learning, models should be **parsimonious**: we seek to reduce model complexity (the number of features required to learn the model, as well as the non-linearity of the function learnt by the model) without sacrificing prediction accuracy. This principle is particularly relevant in this case, as only a limited set of variables can be presented to medical professionals to describe the nature of the model in simple terms, despite its non-linear, "black box" nature. Thus, accurate feature selection and ranking is a critical requirement. Furthermore, we seek to achieve **personalised explanations**, by associating a potentially different list of variables, along with their relative weight, for each individual prediction.

Finally, when assessing model accuracy, we need to minimizes the risk of under-estimating the severity of a patient's condition, that is, by reducing false negatives possibly at the expense of an increase in the number of false positives.

Our modelling approach, which takes account of all of these requirements, is presented in Sec. 3.

1.2 Case study

In addition to testing theoretical model performance as reported in Sec. 3, we have empirically validated our model on an early patient, a 55 year old man who was initially admitted to the clinic with typical COVID pneumonia symptoms. He was treated and discharged the next day after his clinical assessment established a stable condition, with our outcome variable PaO_2/FiO_2 = 420 mmHg well above the 250mmHg cutoff. However, he was readmitted to hospital four days later with severe symptoms, and at that point his condition has worsened to PaO_2/FiO_2 = 230 mmHg.

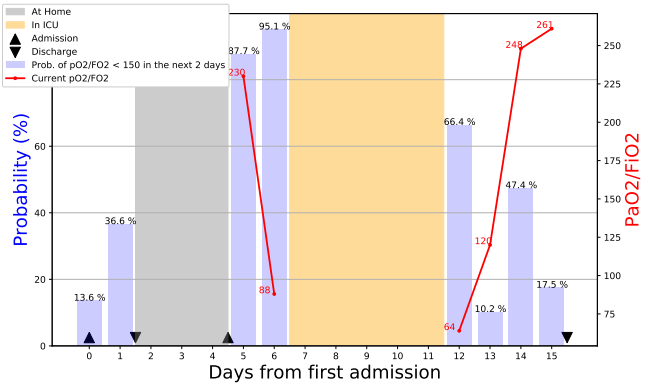

Figure 1. Case study: a patient's journey with retrospective model predictions

In the following 24 hours, the patient experienced a clinically unpredictable dramatic worsening (PaO_2/FiO_2 = 88 mmHg, respiratory rate higher than 35 breaths per minute). He was then transferred to ICU where non-invasive mechanical ventilation (NIV) was started. After 8 days of assisted spontaneous breathing, he was weaned from NIV and discharged the following day without oxygen supply.

We retrospectively used this patient's baseline assessment and then his repeated medical evaluations to predict the probability of adverse outcome at different points in time. The model predicted a 36% probability based on the first admission, followed by much higher confidence after the second admission and prior to ICU treatment, as shown in Fig. 1. If the model have been available at that time, it would have correctly alerted clinical staff against complacence with the first dismissal, which was objectively justified at the time but could not account for the population data that was instead available to the model.

We conclude that, in this particular instance, an early deployment of our model would have provided effective support to assist clinical judgment.

1.3 Contributions

We present our experience in developing a machine learning model that satisfies a number of special requirements and addresses the challenges outlined in the previous Section. The model proves that you can bootstrap decision support to clinical staff in a time of respiratory crisis, by making the best of an approximately curated, constantly evolving dataset by rapidly customising out-of-the-box ML algorithms.

Specifically:

- we have adapted LightGBM and designed a bespoke loss function which includes a penalty β, that can be tuned to adjust the model's FNR (on a test dataset).
- we have successfully experimented with SHAP in combination with LighGBM to provide post hoc explanations of the model's prediction, both globally and locally.

The model is currently available through a hospital private web service that lets clinicians probe the model with new cases and get a prediction using a simple Web interface completely integrated with their usual medical records management system.

This is ongoing work, as the models presented in this work are being periodically re-trained as more patients are added to the dataset.

1.4 Related Work

The global spread of the pandemic by COVID-19 has raised the level of attention of researchers around the world with regard to the monitoring, treatment and study of diseases related to it. One of the most important and characterizing aspects of this pathology is the rapid evolution of the patients' health into a crisis, requiring ICU treatment. This makes ICU allocation strategies a priority and a challenge[9, 14, 8].

Our work aims to provide support to medical decisions both in the triage phase and in the continuative patient monitoring phase. Similar studies have been presented very recently [7, 15, 10, 4]. These study share the common goal to segregate the most critical patients, i.e. those requiring ICU or even approaching death, from those who are improving their health status. The approaches generally used take into consideration very similar variables, including biomarkers, symptoms and co-morbidities, while the outcome chosen may differ, and typically includes a critical respiratory event or death.

The value of our work which makes it stand out in this space is primarily its focus on ensuring trust by clinical staff. As we have explained, we achieve this through a combination of a data-driven strategy for feature set reduction and prioritisation, combined with a loss function that ensures conservative predictions that penalise false negatives. Testimony to this focus is the current experimental deployment of the model as part of the hospital's Information Systems, and its upcoming availability throughout the province.

2 Dataset characterisation

The datasets used in this work were extracted from the Hospital Information System of Policlinico di Modena, where a tailored data collection protocol was implemented in order to gather new data which were deemed relevant to assess the health status of COVID-19 patients. While at the beginning of the pandemic the schema for the new data was vey close to that of the existing EHR management system, new elements were increasingly added, as it became clear that new biomarkers and symptoms were going to be relevant.

We performed two milestone data extractions, after 27 and after 37 days from the start of the covid-specific data collection, which included 91 of the available 99 variables. Specifically, we collected a set of static variables, specifically sex and age, and the 14 most relevant co-morbidities such as diabetes, cardiovascular diseases, neoplasms, and hypertension. We also collected a number of time-varying variables, which are measured periodically, as follows:

- 39 blood and urine tests, including standard blood test and COVID-related ones such as Interleukin-6, Lymphocytes, Troponin and C-reactive protein (CRP);
- 7 blood gas analysis (BGA) measures: pH, $PaCO_2$, SO_2, $Lactates$, PaO_2, HCO_3 and FiO_2;
- 29 different disease specific symptoms, e.g. dyspnoea, cough, fever, conjunctivitis and shivers, and signs, e.g. heart rate, body temperature and respiratory rate.

Time-varying information were collected at different times. For instance, most blood tests are collected daily but some specific tests are collected "on-demand" based on clinical needs. An example is Interleukin 6 which is collected only twice a week and only for the patients that were treated with the immune active drug Tocilizumab.

It is worth noting that we were forced to exclude variables related to both drug therapies because they were collected only for the few patients that received specific treatments and then were almost always missing.

A separate training set was derived from each of the two data extractions. As mentioned in the introduction, the selected outcome was a binary variable stating whether the patient would develop ARDS in the next two days, measured as a PaO_2/FiO_2 ratio lower than 150 mmHg. The EHR of each patient p was therefore sliced in daily snapshots and each sample, exemplified in Fig. 2, was represented by the pair $(x_i^p, y_{i+1,i+2}^p)$ where x_i^p is the feature vector of the day i that contains

- the values of the static variables for the patient p;
- the value recorded at day i for each daily-collected variables;
- the last recorded value for each "on demand"-collected variable.

As far as the outcome $y_{i+1,i+2}^p$ is concerned, we considered the two alternative options: the value of $y_{i+1,i+2}^p$ is set to TRUE either when $PaO_2/FiO_2 < 150$ *both* at day $i+1$ and at day $i+2$ (AND condition), or when $PaO_2/FiO_2 < 150$ *at least* one the two days $i+1$ or $i+2$ (OR condition). The two extractions contained the records for 224 and 287 patients and 2454 and 2888 observations, respectively.

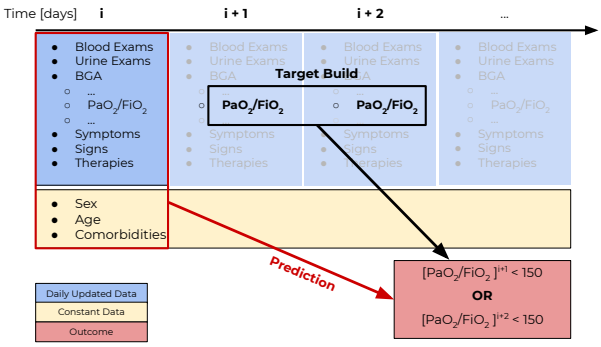

Figure 2. Baseline and follow-up variables and predicted outcome.

Data sparsity and balancing issues for each of the two datasets are summarised in Table 1, showing the mean and the variance of the percentage of completeness for all the 91 variable and the population distribution with respect to the two alternative outcomes, AND conditions and OR conditions.

The table reveals several problems having to do with missing data. With only sex and age complete, the averages are predictably low. Importantly, the number of collected values for some important variables were very low. For instance 168 values were available for lymphocytes (7.5%) and 515 for Interleukin-6 (20%) in the first extraction. This increased slightly, to 557 (7.8%) and 642 (22.2%), respectively, in the second extraction. Furthermore, the completeness of the "on-demand" variables generally decreased in the second batch because these variables are missing for most of the snapshots added in the second batch. As a consequence, the mean percentage of completeness decreased from 62% to 57%.

		Total	Population Distribution FALSE	TRUE	Completeness
First Data Extraction		2454			62% ± 22
	OR Condition	974	417 (43%)	557 (57%)	
	AND Condition	706	419 (59%)	287 (41%)	
Second Data Extraction		2888			57% ± 22
	OR Condition	1068	465 (44%)	603 (56%)	
	AND Condition	796	479 (60%)	317 (40%)	

Table 1. Population report of the two consecutive database extractions.

Sparsity represents a problem because most off-the-shelf learning algorithms do not tolerate missing data. At the same time, removing variables or data points was not an option, as sparsity affects most variables and the training set is of very limited size. Imputation is not an option, either, as the interesting signal about some of the important variables is actually their abrupt changes within a patient's timeline.

On the other hand, the data in Table 1 suggests that using the "OR" outcome, leads to a larger training set, because it only requires PaO_2 and FiO_2 to be available on any single day. This is in contrast to the "AND" condition, where two consecutive data points are required. For instance, considering the most recent data extraction, the outcome built on the OR condition is available for 198 patients (68.9%) and 1068 snapshots (36.9%) that actually represents the potential number of samples. The distribution between the two classes, FALSE and TRUE, is quite balanced, i.e. 417 (43%) and 557 (57%) samples, respectively. In contrast, the distribution for the outcome built on the AND condition is available only for 165 patients (57.4%) and 796 snapshots (27.5%) distributed between the two classes in 419 (59%) and 287 (41%) samples respectively.

Based on these considerations, we therefore decided to use the OR condition, which is at the same time clinically valid, and suitable from a ML modelling point of view, providing a larger and more balanced training set, with a better representation of the positive class.

3 Modelling approach and results

Our main requirements for modelling include: (i) robustness to dataset evolution, sparsity, and possible class unbalance; (ii) feature selection (parsimony); and (iii) control over False Negatives. Here we present our approach to modelling that meets these requirements.

3.1 Robustness

Robustness is achieved mainly through a choice of a learning algorithm that can tolerate missing data, and the appropriate tuning of the hyper-parameters. For this binary classification problem the main candidates are XGBoost[6] and LightGBM[7]. Both implement a form of decision tree ensembles that can tolerate missing data in a tunable way, where imputation is not an option as explained above. Decision tree algorithms are based on the idea that at each node, the feature is chosen to maximise some measure of information gain. XGBoost extend the idea to missing values, by assigning a *default direction* to each node in the tree, using a technique known as *Sparsity-aware split finding* [6]. This determines which way the decision proceeds when a feature value is missing and the corresponding node condition cannot be evaluated. While both algorithms can also be retrofitted with explanations, as described below. LightGBM was selected owing to its greater flexibility.

3.2 Model and Feature selection

Following standard experimental practice, the choice of learning algorithm and features used to learn the model go hand in hand. Our process involved (i) selecting a category of features with sufficient predictive power, and (ii) reducing the feature space within that class, in order to facilitate model interpretability. For this learning problem we compared three models, which were built using the three sets of features described earlier, namely (i) the entire set of 91 variables in the dataset, (ii) only the 39 biomarker variables; and (iii) the 31 "symptoms and signs" variables. The results are reported in Table 2.

The first model yields the best performance, however it was unlikely it would have been useful in practice as it required extensive and expensive data collection for each patient. It was also feature-rich and possibly redundant. The simpler subset of biomarkers, used for the second model, is attractive as the data acquisition workflow is entirely standard, and it provides an objective assessment of the patient's health status. Its performance is comparable with that of the first, with a slight advantage in the number of FN.

Finally, the 31-variables model was appealing in terms of data collection, as the "signs and symptoms" variables included only questions to patients and simple measurements such as heart rate and temperature, which are easy to obtain. However it exhibited inferior performance (AUC=0.69) along with a higher number of both FN and FP relative to the previous two models. This result suggest that such subjective data is quite less informative than objective and instrumental data collection.

All models were generated using standard ML practice, namely a $75/25$ training / test split, random selection from the majority class for balancing, 10-fold cross-validation, and hyper-parameter tuning session using Grid Search.

For feature set reduction we adopted a data-driven approach centred on Shapley values[8]. These are generated by a values-based ML model interpretation framework that provides both a global-level assessment of the relative importance of each feature used by the model, as well as a local view of how each feature is weighted when making individual predictions. Furthermore, Shapley values offer an interpretation of feature importance as a function of the value of the feature. This leads to an intuitive interpretation, for instance "high values of Dyspnea contribute strongly to an adverse outcome, while low values make the feature relatively less important". This sort of explanation provides clinicians with a way to validate the model against their own expertise, and thus may contribute to building trust

[6] `https://github.com/dmlc/xgboost`
[7] `https://github.com/microsoft/LightGBM`
[8] `https://github.com/slundberg/shap`

Dataset	TN	FP	FN	TP	Acc.	Prec.	Rec.	F1	AUC
91 Mixed variables	95	21	40	111	0.77	0.84	0.74	0.78	0.85
39 Biomarkers only	89	27	39	112	0.75	0.81	0.74	0.77	0.83
31 Signs and Symptoms	74	42	56	95	0.63	0.69	0.63	0.66	0.69

Table 2. Performance of the three models based on the dataset choice

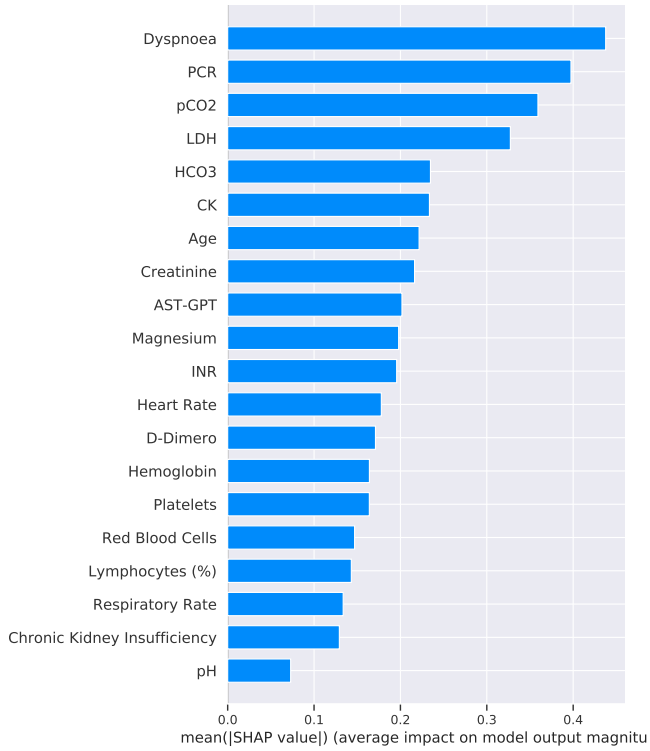

Figure 3. The 20 final variables in order of importance according to the SHAP model explainer

3.3 Controlling False Negatives

The model emerging from the two steps described above can be further tuned to achieve a trade-off between False Negatives and accuracy. This was achieved by adding a new hyper-parameter β to the loss function that was minimised during the learning process. The customized loss-function is therefore defined as follows:

$$L(y, p(x)) = -\beta \cdot y \ln p(x) - (1 - y) \ln (1 - p(x)) \quad (1)$$

When $\beta = 1$, we get the standard Log-loss function. The effect of tuning β when training the 20-variables version of the model is reported in Table 3. The table shows a clear trade-off between FN and FP, as expected. The obvious measure that balances the two is FP/FN, which is closest to 1 for $\beta = 2$. As this is also the setting where accuracy begins to drop, it is the one we used in the rest of the experiments. Moreover, note that the FNR of this model is even lower than in the earlier 39-variables model. The ROC curve of this final model is depicted in figure 4.

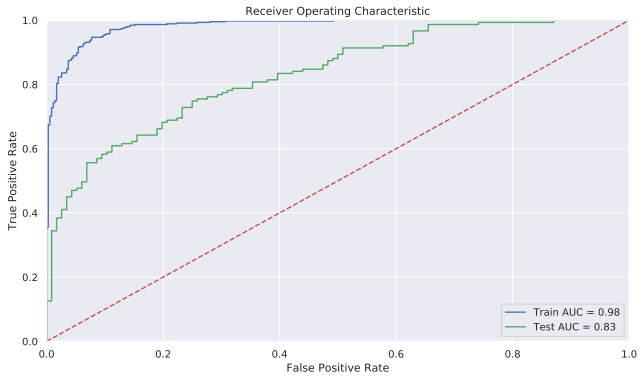

Figure 4. AUC/ROC curve of the model trained with the 20 variables dataset and the custom loss function with $\beta = 2$

in the model's predictions. Also, the framework uses a "post hoc" method for generating such explanations, which can be used in conjunction with non-linear models such as the decision trees generated by LightGBM (for a linear model, feature ranking is simply given by the model's parameters).

Using this approach and global-level feature rankings, and starting from the original 91 variables, we repeatedly pruned features from the set and retrained the model with the remaining features, using a combination of performance and FNR to guide the process. The final set consists of 20 core features, shown in Fig. 3. Note that these are drawn from each of the variable categories: biomarkers, BGA, signs, symptoms and co-morbidities, confirming that multiple views on a patient's status are required to generate reliable predictions.

The first row of Table 3 shows the performance of the resulting 20-variables model 3.3.

4 Concluding remarks and future works

The ML model presented in this paper is currently deployed behind a web service and integrated into the Hospital's Information Systems where it can be used for medical decision support systems in two phases of the patient assessment. Firstly, as a first triage evaluation, e.g. in the Emergency Room, and secondly after admission and through patient's monitoring during their stay in hospital.

An ongoing challenge with this project is the rapidly moving dataset, both in schema as well as in content, as pointed out through the paper. This translates into rapid changes in data distribution, sparsity, and balance with respect to the outcome, and thus into the need to periodically re-configure and re-train the model, including potentially changing the implementation when it is no longer suitable.

β	TN	FP	FN	TP	Acc.	Prec.	Rec.	F1	AUC
1 (the binary log-loss)	87	29	40	111	0.74	0.79	0.74	0.76	0.84
2	82	34	35	116	0.74	0.77	0.77	0.77	0.83
3	71	45	29	122	0.71	0.71	0.79	0.75	0.82
4	68	48	25	126	0.73	0.72	0.83	0.78	0.82
5	63	53	18	133	0.73	0.72	0.88	0.79	0.82
6	56	60	15	136	0.72	0.69	0.90	0.78	0.82

Table 3. Performance of the 20 variables model as a function of the penalty parameter β

This situation is also ideal for experimenting with "AutoML" approaches, which are becoming increasingly popular both in the scientific community and as part of commercial offering. These solutions aims to build automatically tailored ML pipelines and include operators for data transformation such as data imputation, feature processing, classification and calibration algorithms. They have proven to be of remarkable reliability and performance. One notable example is AutoPrognosis [2], specifically tailored to clinical ML pipelines and uses ML itself to proper configure and chose the best configuration for the ML models it's creating.

Adopting AutoML solutions is planned for future work, in the hope to improve the efficiency of the whole ML pipeline design, and also to test such approaches in new emergency contexts, in which there is no time for the manual development of the pipeline itself (COVID-19 pandemic would have been an example).

ML systems which are also capable of explaining the behavior of the model can go far beyond mere prediction. Model interpretation approaches in medical applications can be extremely useful, for example, to discover new predictors for the chosen outcome. Also, with the appropriate choice of population it may also be possible to identify different predictors for different sub-populations, leading to new insights into data-driven personalisation of predictive models.

These approaches are becoming more and more effective and popular in medical applications [3, 1, 5], also in the recent COVID-19 pandemic context [7, 10, 15]. As mentioned, our approach to explanatory models involves Shapley values [12, 11], which provide a measure of the impact of each variable on the construction of the predictions result for every instance. More specifically, these are positive or negative numerical values and, in a binary classification task, the sign indicates whether the variable contributes to the positive or to the negative class. This provides an immediate and perception of which of the variables have a positive vs negative impact on the patient's outcome.

For instance, in our models it appears that lab variables are stronger predictors of a patient's clinical condition. however, Shapley values show that these are best combined with non laboratory variables when it comes to explaining the patient's outcome and obtaining better personalized insights on the individual. This can be seen in the global view of these variables, shown in Fig. 3, and in Fig. 5, where we show the local interpretation of the case study patient data on the most critical day we have from his medical records (95% probability of having respiratory failure within the next 2 days predicted exactly the day before being transferred to ICU under mechanical ventilation). Two of the most important variables here, respiratory rate and dyspnoea, are not laboratory variables. This indicates how some of these elements can be relevant to the analysis of health status and in the formulation of a therapeutic plan.

Finally, an interesting further study is to extend the model to forecasting the entire patient's journey through stages of disease and treatment while in hospital, from admission to discharge. Understanding the evolution of the patient's state can be of great importance both in the personalisation of the care plan, in the prevention of adverse outcomes, as well as in the planning of hospital resource organization.

REFERENCES

[1] Ahmed M. Alaa, Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, and Mihaela van der Schaar, 'Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants', *PLOS ONE*, **14**(5), 1–17, (05 2019).

[2] Ahmed M. Alaa and Mihaela van der Schaar, 'Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning', *CoRR*, **abs/1802.07207**, (2018).

[3] Ahmed M. Alaa and Mihaela van der Schaar, 'Prognostication and risk factors for cystic fibrosis via automated machine learning', *Scientific Reports*, **8**(1), 11242, (Jul 2018).

[4] Egon Burian, Friederike Jungmann, Georgios A. Kaissis, Fabian K. Lohöfer, Christoph D. Spinner, Tobias Lahmer, Matthias Treiber, Michael Dommasch, Gerhard Schneider, Fabian Geisler, Wolfgang Huber, Ulrike Protzer, Roland M. Schmid, Markus Schwaiger, Marcus R. Makowski, and Rickmer F. Braren, 'Intensive care risk estimation in covid-19 pneumonia based on clinical and imaging parameters: Experiences from the munich cohort', *Journal of Clinical Medicine*, **9**(5), (2020).

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, 'Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, p. 1721–1730, New York, NY, USA, (2015). Association for Computing Machinery.

[6] Tianqi Chen and Carlos Guestrin, 'Xgboost: A scalable tree boosting system', in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, (2016).

[7] Yuanfang Chen, Liu Ouyang, Sheng Bao, Qian Li, Lei Han, Hengdong Zhang, Baoli Zhu, Ming Xu, Jie Liu, Yaorong Ge, and Shi Chen, 'An interpretable machine learning framework for accurate severe vs non-severe covid-19 clinical type classification', *medRxiv*, (2020).

[8] Ezekiel J. Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P. Phillips, 'Allocating medical resources in the time of covid-19', *New England Journal of Medicine*, **382**(22), e79, (2020).

[9] Ezekiel J. Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P. Phillips, 'Fair allocation of scarce medical resources in the time of covid-19', *New England Journal of Medicine*, **382**(21), 2049–2055, (2020).

[10] Frank Stefan Heldt, Marcela P Vizcaychipi, Sophie Peacock, Mattia Cinelli, Lachlan McLachlan, Fernando Andreotti, Stojan Jovanovic, Robert Durichen, Nadezda Lipunova, Robert A Fletcher, and Anne *et al* Hancock, 'Early risk assessment for covid-19 patients from emergency department data using machine learning', *medRxiv*, (2020).

[11] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee, 'Explainable ai for trees: From local explanations to global understanding', *arXiv preprint arXiv:1905.04610*, (2019).

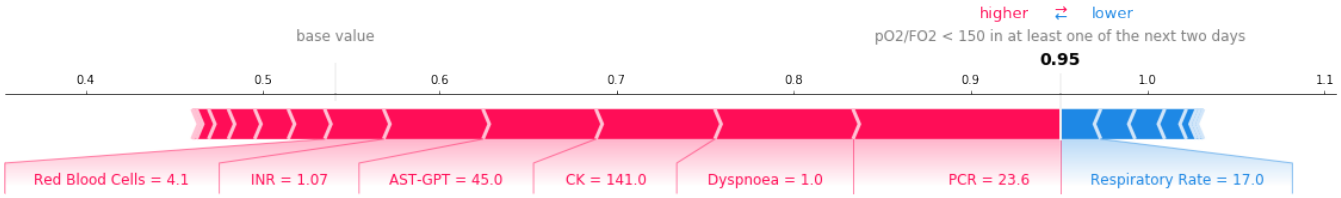

Figure 5. Local Interpretation for the 6th and most critical day for the patient of the case study in section 1.2

[12] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774, Curran Associates, Inc., (2017).

[13] Lingzhong Meng, Haibo Qiu, Li Wan, Yuhang Ai, Zhanggang Xue, Qulian Guo, Ranjit Deshpande, Lina Zhang, Jie Meng, Chuanyao Tong, Hong Liu, and Lize Xiong, 'Intubation and Ventilation amid the COVID-19 Outbreak: Wuhan's Experience.', *Anesthesiology*, **132**(6), 1317–1332, (jun 2020).

[14] Robert D. Truog, Christine Mitchell, and George Q. Daley, 'The toughest triage — allocating ventilators in a pandemic', *New England Journal of Medicine*, **382**(21), 1973–1975, (2020).

[15] Akhil Vaid, Sulaiman Somani, Adam J Russak, Jessica K De Freitas, and Fayzan F *et al.* Chaudhry, 'Machine learning to predict mortality and critical events in covid-19 positive new york city patients', *medRxiv*, (2020).