Understanding and Extrapolating from Healthcare Data through Machine Learning: a Case Study on a COVID-19 Dataset

Michele Rispoli[1,*], Francesco Salton[2,3], Andrea Rocca[2], Paola Confalonieri[2,3], Marco Confalonieri[2,3], Alberto D'Onofrio[1] and Luca Manzoni[1,*]

[1] Department of Mathematics, University of Trieste, Informatics and Geosciences, Trieste, Italy
[2] Department of Medical Surgical and Health Sciences, University of Trieste, Trieste, Italy
[3] Pulmonology Unit, University Hospital of Cattinara, Trieste, Italy

Abstract

Machine Learning (ML) and Artifical Intelligence (AI) are increasingly more present in every aspect of research and industry, aiding (and some times replacing) humans in understanding and interacting with the world. In healthcare, these models can assist practitioners in interpreting data, providing useful insights and enabling the prediction of future therapy outcomes. In this short paper we review two retrospective studies that we conducted on a dataset of COVID-19 patients, presenting concrete examples of how these technologies can be applied in health research.

Keywords

Machine Learning, Classification, Unsupervised Learning, Survival Analysis, COVID-19,

1. Introduction

Machine learning (ML) and **artificial intelligence (AI)** have gradually become a standard in both research and industry: thanks to the increasingly more capable models developed in these fields, we are more than ever able to automate tasks and distill crucial information from data, both in the public and private sectors, with notable consequences on the whole society. The perspective of a future in which most tasks are accomplished in collaboration with (or completely autonomously by) AI agents is becoming increasingly more realistic. This urges us to focus our attention on the role of us humans in this context, and in particular, on how we process information, distill knowledge, and extrapolate from past experience to orient our future actions: such reflections are crucial for understanding how we can leverage these technologies to aid our efforts, guiding how we design and use ML algorithms and AI systems, taking into account the appropriate technical, practical and ethical aspects.

Focusing on the main subject of this work, healthcare is indeed one of the fields in which the advent of AI has already brought numerous benefits for the whole community, as it can help practitioners in interpreting clinical data, predicting the outcome of therapies, and generally improve the efficacy of medical treatments [1, 2].

In this work we will briefly review two studies that we conducted on a specific health dataset, providing two concrete examples of how ML may be used in practice to extract useful knowledge from medical data.

We will begin with an overview of the dataset (Section 2), and then provide a brief summary of the studies that we conducted on it (Sections 3 and 4).

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*] Corresponding author.

michele.rispoli@phd.units.it (M. Rispoli); alberto.donofrio@units.it (A. D'Onofrio); lmanzoni@units.it (L. Manzoni)
0009-0005-7350-1628 (M. Rispoli)

2. COVID-19 Dataset

The raw dataset comprises information on 1012 patients that were hospitalized between February 2020 and May 2022 due to COVID-19 related pneumonia, which participated in two formerly published multicentre studies in northern italy investigating the efficacy of different gluco-corticoid (GC) treatments [3, 4]. Extensive cleaning, error correction and homogenization was necessary to make the data effectively usable, finally resulting in a pool of **951 patients and 87 features**. Available features include demographic information, medical history, therapy, complications, relevant dates, outcome, and up to five serial measurements for each of two physiological quantities, namely: the arterial partial pressure of oxygen to fraction of inspired oxygen ratio (PaO2 /FiO2 , mmHg), and the C-reactive protein levels (CRP, mg/L).

From a data-analysis perspective, this dataset presents several challenges: fortunately for the patients, the registered decease rate is relatively limited (slightly less 15%), which translates to high **unbalancedness** w.r.t. to the outcome; even after extensive cleaning, missing values are present in most features, and data is highly inhomogeneous: features comprise a **mixture of numerical, categorical, and (most prominently) binary features**, and also include **short time series** (complete with the associated time data features), thus demanding more elaborate pre-processing; finally, while the dataset's size is above average, if compared with sample sizes of similar retrospective studies in health research, it falls short of fulfilling the requirements to apply and evaluate more data-intensive techniques (e.g., deep neural networks), especially in combination with the other limitations.

3. Study 1: Predicting in-hospital mortality

In our first study, we developed a machine learning algorithm to **predict in-hospital mortality of COVID-19 patients treated with GCs within 30 days from hospitalization** [5] (that is, we tackled an instance of binary classification problem with a supervised approach).

Patients which weren't treated with GCs were excluded, and an initial feature pre-selection and extraction phase was performed, to limit the number of undefined values and to ensure an homogeneous representation for each patient. The final data pool for this study comprised **825 patients and 52 features**.

We developed a pipeline to train six machine learning algorithms for this prediction task, and evaluate their performance, both during training and onto a left-out 20% portion of the dataset (i.e., internal validation).

The pipeline included a **feature selection** procedure, which leveraged XGBoost to rank the features basing on their contribution to the generalization capabilities of the model (estimated via 5-fold-CV onto the training set). By applying, we identified an **optimal reduced set of 9 features**, namely: CRP improvement, PaO2/FiO2 improvement, age, coronary artery disease, need for invasive mechanical ventilation (IMV), acute renal failure, chronic heart failure, earliest PaO2/FiO2 ratio sample, and body mass index (BMI).

We then trained and fine-tuned via grid-search **six ML algorithms**, namely: logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost, XGB) and a fully-connected feed-forward neural network (NN). When needed, missing values were filled using mean/modal values from the training data, while performance evaluation relied upon several metrics, most notably AUROC and AUPRC.

Results show that **Random Forest and XGBoost** were the most successful models, scoring a test AUROC and AUPRC of 0.938 (bootstrap intervals 0.903–0.969) and 0.714 (0.548–0.856), and 0.937 (0.901–0.968) and 0.701 (0.538–0.846) respectively. Furthermore, we computed the **Shapley additive explanation values (SHAP)** [6] for the best models, which allowed us to get a better insight on the relation between predictions and input features: according to our model, the most influential features are the presence of an **improvement of the PaO2/FiO2 ratio** (absolute mean SHAP: 0.149), **Age** (0.13) and **improvement of CRP levels** (0.071), while minor protective effects (<0.03) resulted associated with the absence of IMV treatment and absence of previous history of acute renal failure, coronary artery disease and chronic heart failure.

4. Study 2: Identifying and characterizing clusters with different survival behaviors

In our most recent study, which is still in the workings, we investigate the possibility to model the survival trajectory of the patients in our data, this time also taking into account when information becomes available during hospitalization, and adopting an unsupervised learning approach. More specifically, we wish to **cluster the population basing on the values of both static and time-varying features**, focusing on a predetermined series of key **landmark times**, and then perform **survival analysis** [7] on these clusters, in order to determine (i) if and when the available features present geometric patterns that can be exploited by **spectral clustering** [8, 9] to partition the dataset, (ii) whether the resulting sub-populations differ in terms of survival, and, finally (iii) how well can we distinguish these subpopulations, when only considering one variable at a time.

In fact, the primary goal of this study is presenting our novel technique that combines unsupervised clustering with landmark survival analysis, which is **well suited for datasets of limited size**, and could thus find application in several real-world research settings, not limited to healthcare; in this context, the analysis of our dataset mainly serves demonstration purposes.

The data preparation phase for this study required the identification (or extraction) of time-varying features, as well as the definition of a strategy to evaluate them at different times, which involved a more extensive review and manipulation of date fields (which were mostly disregarded in our first study). The resulting dataset has **946 patients** (at $t = 0$, that is, at hospitalization) and **24 features** (9 of which are time-dependent, and only become available from the second landmark, at day 2).

Key landmark times were defined at days 2, 4, 8 and 15 from hospitalization, basing on the times at which PaO2/FiO2 and CRP values were sampled. Notice that the further we advance in time, the less patients are still hospitalized (patients who die or are dismissed are no longer considered), though, at the same time, more information on them becomes available, as more clinical samples are collected and more therapies are applied.

Our goal, for each of these landmarks, is using spectral clustering to split the cohort into sub-population, basing on the features available at that time, and then estimate and compare their survival curves. To begin, we define a measure of similarity between patients: we chose to equally weight the differences across the available features, privileging the idea that doing otherwise would introduce additional bias. We then determine the number of clusters that we shall use at all landmarks, basing on the population at baseline, which in our case results to be 2. Subsequently, we apply spectral clustering with the chosen parameters at each landmark, furthermore refining the clustering by applying a backwards variable selection scheme, ensuring that variables which are less informative (at that landmark) are excluded from the computation of the clusters. Finally we compute the Kaplan-Meier survival curves for each cluster, and compare them with log-rank tests and in terms of Cox hazard-ratio w.r.t. the clustering labels, furthermore investigating how the two clusters differ in regards to their features' distributions by means of statistical testing.

The the full results will be made available in a dedicated publication, though preliminary results show that our method successfully manages to partition our cohort into **a low-risk and a high-risk group**, with the latter consistently presenting a worse survival curve: the **hazard ratio** between the clusters starts at 1.64 at hospitalization, and reaches a **peak of 5.43 at t=8**; furthermore, uni-variate testing indicate that **history of hypertension** presents the most distinguishable distributions across the clusters (chi-squared test p-value consistently below 10^{-16}) especially at the initial landmarks (p-value $< 10^{-100}$), readily followed by **age** (Kolmogorov-Smirnoff test p-value consistently below 10^{-8}), and, at later landmarks, by features extracted from serial **PaO2/FiO2 ratio and CRP** measurements.

These results demonstrate the efficacy of this technique, which combines the exceptional ability of unsupervised learning in extracting complex geometrical patterns hidden in the data, with the expressivity and simplicity of landmark survival analysis: if such patterns are present in the data and are indicative of differences in survival, our method provides a way to spot and characterize them, identifying when and in which features these differences manifest more prominently. Furthermore, while in this study we use this technique for inference, it could be easily extended for predictions, for example, by training a classifier onto the already clustered data (possibly exploiting the knowledge gained from the final uni-variate analysis).

The results on our data are very promising, nonetheless we acknowledge that our technique isn't guaranteed to always uncover relevant patterns (e.g., clustering may be impossible, or the resulting clusters may not present notable differences in terms of survival and/or single features), as this ultimately rests on the available data. Furthermore, this technique was designed with the intended use of analyzing survival datasets of modest size (i.e., within the thousands of samples), in particular spectral clustering can be particularly cumbersome for large datasets; while it would be possible to apply some modifications to cut the computational costs (e.g., sub-sampling, using clustering algorithms that scale better), in such cases we would rather recommend considering other techniques, such as deep survival analysis [10, 11], which are better suited for processing larger datasets.

5. Conclusions

In this work, we presented two studies that we conducted on a specific healthcare dataset, offering a concrete demonstration of how machine learning techniques can aid practitioners in extracting useful information from data.

In the first study, we developed a classification algorithm for predicting mortality within 30 days for patients that are hospitalized with COVID-19 related pneumonia; the final models achieved satisfactory performances in internal validation, and the analysis of its SHAP values allowed to determine that the improvement of PaO2/FiO2 ratio and CRP were the most influential in determining the predictions.

In our second study (which is still in the workings), we combine unsupervised spectral clustering with landmark survival analysis, to identify and characterize subgroups of the population which experience different levels of mortality risk. Preliminary results indicate that our cohort can be effectively split in two risk groups (high and low), with the greatest separation between the respective survival curves occurring at day 8 from hospitalization, furthermore identifying history of hypertension, age, and features related to PaO2/FiO2 ratio and CRP as most distinctly distributed characteristics between the two clusters, with varying degrees of separation across the chosen landmark times.

Both studies are examples of how, with the help of ML, complex medical data can be extrapolated to aid future decisions, and distilled into a form that is easier to understand for the human practitioners. We believe that such efforts are and will remain relevant: providing alternative ways to interpret and view data constitutes the very basis of human understanding, and we believe it is fair to recognize this to be true also about AI systems, which increasingly more often help (or even replace) us in the research of knowledge.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Holmes, L. Sacchi, R. Bellazzi, et al., Artificial intelligence in medicine, Ann R Coll Surg Engl 86 (2004) 334–8.

[2] R. Perrault, J. Clark, Artificial intelligence index report 2024 (2024).

[3] F. Salton, P. Confalonieri, G. U. Meduri, P. Santus, S. Harari, R. Scala, S. Lanini, V. Vertui, T. Oggionni, A. Caminati, et al., Prolonged low-dose methylprednisolone in patients with severe covid-

19 pneumonia, in: Open forum infectious diseases, volume 7, Oxford University Press US, 2020, p. ofaa421.

[4] F. Salton, P. Confalonieri, S. Centanni, M. Mondoni, N. Petrosillo, P. Bonfanti, G. Lapadula, D. Lacedonia, A. Voza, N. Carpenè, et al., Prolonged higher dose methylprednisolone versus conventional dexamethasone in covid-19 pneumonia: a randomised controlled trial (medeas), European Respiratory Journal 61 (2023).

[5] F. Salton, M. Rispoli, P. Confalonieri, A. De Nes, E. Spagnol, A. Salotti, B. Ruaro, S. Harari, A. Rocca, A. d'Onofrio, et al., A tailored machine learning approach for mortality prediction in severe COVID-19 treated with glucocorticoids, The International Journal of Tuberculosis and Lung Disease 28 (2024) 439–445.

[6] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[7] J. P. Klein, H. C. Van Houwelingen, J. G. Ibrahim, T. H. Scheike, Handbook of survival analysis, CRC Press, 2014.

[8] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (2007) 395–416.

[9] L. Ding, C. Li, D. Jin, S. Ding, Survey of spectral clustering based on graph theory, Pattern Recognition (2024) 110366.

[10] R. Ranganath, A. Perotte, N. Elhadad, D. Blei, Deep survival analysis, in: Machine Learning for Healthcare Conference, PMLR, 2016, pp. 101–114.

[11] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, A. Bender, Deep learning for survival analysis: a review, Artificial Intelligence Review 57 (2024) 65.