UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVIII CICLO DEL DOTTORATO DI RICERCA IN

APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Fondazione Cassa Risparmio di Trieste

Machine Learning Methods for Clinical Decision Support: an Analysis Based on COVID-19 Data

Settore scientifico-disciplinare: INF/01

DOTTORANDO
MICHELE RISPOLI

COORDINATORE
PROF. FRANCESCO PAULI

SUPERVISORE DI TESI
PROF. LUCA MANZONI

CO-SUPERVISORE DI TESI
PROF. ALBERTO D'ONOFRIO

ANNO ACCADEMICO 2024/2025

Università degli Studi di Trieste

Dipartmento di Matematica, Informatica e Geoscienze

PhD THESIS IN APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Machine Learning Methods for Clinical Decision Support: an analysis based on COVID-19 Data

PhD Candidate:
Michele Rispoli

Supervisor:
Prof. Luca Manzoni

Co-supervisor:
Prof. Alberto d'Onofrio

December, 2025

Acknowledgements

I would like to thank my PhD supervisors Luca Manzoni and Alberto d'Onofrio, whose guidance and expertise supported me in my research path, and were fundamental for the development of the work presented in this thesis.

My PhD was funded by Fondazione Cassa Risparmio di Trieste, whose support is gratefully acknowledged.

I would also like to thank Marco Confalonieri, Francesco Salton, Andrea Rocca, and all the other collaborators from the pneumology unit of the University Hospital of Cattinara, Trieste, who provided valuable field expertise and access to the data which made this work possible.

Trieste, Italy
December, 2025

Contents

Abstract

Between the years 2020 and 2023, **COVID-19** posed an unprecedented challenge to healthcare systems worldwide, rapidly evolving into a pandemic that claimed millions of lives. While the disease has now reached an endemic stage, it continues to demand clinical attention, and the prospect of future pandemics remains a concrete threat. Consequently, developing robust data-driven tools to support healthcare and emergency response remains of utmost importance.

In this context, **Machine Learning (ML)** and **Artificial Intelligence (AI)** have proven to be valuable allies for healthcare professionals, enabling the extraction of meaningful insights from clinical datasets, accelerating workflows, and supporting personalized care in both diagnostic and prognostic settings.

This thesis contributes to these ongoing efforts by developing and applying novel **ML techniques for the analysis of clinical tabular data**, with a particular focus on COVID-19. The research was conducted in collaboration with the pneumology unit of the University Hospital of Cattinara, Trieste, which is part of *ASUGI* (Azienda Sanitaria Universitaria Giulioano-Isontina), the Public Health Authority for the provinces of Gorizia and Trieste.

Three main studies are presented, each addressing a dual objective:

1. to design and validate methods for analyzing tabular clinical datasets, thereby providing ML practitioners with new methodological tools; and
2. to apply these methods to a real-world COVID-19 dataset to derive actionable insights for clinical decision-making.

The first study presents a comprehensive ML pipeline to **predict in-hospital mortality** of patients with severe COVID-19 pneumonia treated with glucocorticoids. Six supervised algorithms, ranging from logistic regression and decision trees to ensemble methods and neural networks, were trained and evaluated. The best models achieved strong predictive performance (AUROC > 0.9), demonstrating that accurate predictions can be obtained even from moderately sized datasets through careful preprocessing, feature selection, and validation. Model explainability was ensured using SHAP values, which provided individualized explanations and confirmed the clinical relevance of factors such as age, comorbidities, C-reactive protein, and improvements in the PaO_2/FiO_2ratio.

The second study focuses on **fairness and bias detection in tabular datasets**, presenting an improved version of FanFAIR, a hybrid statistical-ML tool to quantify bias

present in the data. The updated tool supports a wider range of data types, integrates with pandas dataframes, and enables the specification of sensitive attributes. Application of FanFAIR to our COVID-19 dataset revealed that suitable pre-processing can simultaneously enhance model accuracy and fairness, highlighting the importance of methodological and ethical rigor in healthcare AI.

Finally, the third study introduces a novel framework **combining spectral clustering with landmark survival analysis** to identify latent patient subgroups characterized by distinct survival behaviors. Applied to our COVID-19 dataset, the method uncovered clinically meaningful clusters corresponding to high- and low-risk patient groups, whose survival trajectories and clinical profiles remained distinct across multiple temporal landmarks. This approach extends the applicability of ML to survival data analysis in heterogeneous datasets of limited size, where conventional or deep learning models may struggle.

The results presented in this thesis demonstrate the potential of transparent, ethical, and interpretable ML to support data-driven decision making in healthcare, particularly in the context of future epidemic and pandemic preparedness. Furthermore, the developed techniques form a comprehensive methodological toolkit adaptable to similar ML and AI problems, with potential applications extending beyond healthcare to other scientific and industrial domains.

1 Introduction

In the age of data, that we're currently living in, Machine Learning (ML) and Artificial Intelligence (AI) have emerged as tools of choice for processing the torrent of data produced by our society, finding application in virtually any field of knowledge, and, increasingly more often, directly impacting human well-being. In healthcare, ML and AI are allowing clinicians and researchers to mine useful knowledge from electronic health records (EHR), assisting in diagnosing and treating diseases, expediting clinical workflows, boosting pharmaceutical research, and generally improving healthcare delivery [1]. While methods from classical statistics are typically employed for describing general properties at population levels, AI and ML can exploit complex patterns in the data to make predictions for individual patients, offering clinical decision support for personalized care [2, 3]. In practice, developing models that fulfill these promising expectations can be very challenging: data quality and volume must be sufficient, which seldom corresponds to the state of raw clinical data, due to the intrinsic complexity of the data and of the procedures to collect it, thus requiring extensive pre-processing [4, 5]. The sensible nature of the clinical setting introduces privacy restrictions that often limit data availability, and imposes the need for AI and ML models that are not only accurate, but also understandable from the clinical personnel who has to take the final decisions. In this context, non-generative ML approaches remain preferable over the more recent generative ones, as they are relatively less data-intensive, and can rely on well-established performance measures, as well as techniques to explain how outputs are computed [6].

The work presented in this thesis focuses on the development of non-generative AI and ML techniques aimed at supporting decision making in a clinical setting. The proposed techniques, while implemented with the primary goal of analyzing a specific dataset of

COVID-19 patients, can be applied for any tabular dataset presenting similar characteristics (i.e. mixed data-types, presence of time-varying features) and challenges (i.e. modest size, unbalancedness of the output classes, data missingness).

The contents of this thesis are based on the following works:

- "A tailored machine learning approach for mortality prediction in severe COVID-19 treated with glucocorticoids", journal paper published on The International Journal of Tuberculosis and Lung Disease 28.9 (2024): 439-445, of which I am co-first author [7];
- "Investigating Fairness with FanFAIR: is Pre-Processing Useful Only for Performances?", conference paper published on the proceedings of 2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM), IEEE, 2025, of which I am first author [8];
- "Spectral Clustering-Powered Survival Analysis for heterogeneous clinical datasets: a case study on COVID-19", submitted manuscript of which I am first and corresponding author [9].

1.1 Context: the COVID-19 pandemic

The COVID-19 (Corona virus disease - 2019) pandemic was a major health crisis, which affected the world globally between the years 2020 and 2023, infecting hundreds of millions and causing millions of deaths worldwide, posing an unprecedented challenge for healthcare systems across the globe. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) - the virus responsible for COVID-19 - was first identified in Wuhan, Hubei province, China in late December 2019 [10], and is believed to have zootic origins [11, 12, 13].

The majority of infected patients develops flu-like symptoms (fever, coughing, shortness of breath) ranging from mild to severe within the first two weeks, although a considerable proportion of the population is asymptomatic carriers, greatly complicating the efforts to track and limit the virus spread. Transmission mainly occurs through the respiratory route with high efficacy, especially considering the elevated reproduction rate of the virus, and its ability to survive on inanimate surfaces [11]. Combined with the ever increasing level of global mobility, these factors favored the rapid spread of the disease during the pandemic onset, inducing governments to apply unprecedented mobility restrictions between and within countries [14].

Italy, in particular, was the first European country to be severely affected by COVID-19: [15] three weeks after the initial outbreaks, registered on the 21st of February in the northern regions of Lombardy and Veneto, the virus had already infected more than 10'000, killing more than 600, and prompting the Italian government to impose a national lockdown. By the 20th of March, Italy had reportedly suffered the highest toll of COVID-19 deaths than any other country in the world, with over 3'400 victims, and by the

beginning of April, the Lombardy region alone reported more than 10'000 deaths - roughly three times the amount reported in China at the time. The exact causes behind this tragic escalation in Italy are hard to pinpoint, as multiple factors might have contributed: reduced personnel due to cuts to the public healthcare system, high median age of the susceptible population, absence of a protocol for testing healthcare professionals, lack of personal protection equipment, a general underestimation of the danger by the both the government and the public, which delayed the enforcement of social containment measures [15]. Nevertheless, Italy was merely one of the first countries to suffer severe consequences, inadvertently pioneering the application of containment measures which most other countries in the world were soon forced to adopt.

The COVID-19 pandemic impacted public health well beyond its mortality rate, disrupting social activities, causing a surge in emergence of mental health issues, and aggravating pre-existing social and economic inequalities. At the same time, the emergency motivated the global research efforts which allowed to develop vaccines and identify treatments such as monoclonal antibodies, antiviral drugs and corticosteroids. Although the pandemic is over, COVID-19 has reached the endemic stage, and still circulates to this day, constituting an ongoing challenge for healthcare systems across the world. Furthermore, never before in history there was such large availability of data regarding the same epidemic event, offering an unprecedented opportunity for ML and AI researchers. Nonetheless, research often had to be carried on datasets presenting several challenges, prompting ML and AI practitioners to develop techniques capable of overcoming them.

1.2 Contributions overview

The research contributions presented in this thesis consist of three studies - two published [7, 8] and one submitted [9]. Each of them had a two-fold objective:

1. **developing techniques to conduct analyses on (clinical) tabular datasets**, providing ML and AI practitioners with novel tools to tackle similar problems (i.e. classification, dataset evaluation, and survival analysis);

2. applying these techniques to **analyze a specific COVID-19 dataset**, obtaining insights which may support clinical decision making in the treatment of this disease.

More specifically, in the first study (i.e. the published journal paper [7]) we propose an original pipeline to train and evaluate six different ML algorithms - namely, logistic regression, decision tree, random forest, extreme gradient boosting, support vector machine and a (shallow) neural network - for **predicting mortality of hospitalized COVID-19 patients**, that is, tackling a binary classification problem in a supervised setting, furthermore employing additive SHAP values, an explainable AI (XAI) technique, to provide detailed **explanation for each individual outcome** predicted by the best models.
In the second study (i.e. the published proceedings paper [8]) we present an improved

version of FanFAIR, a tool which leverages both classical statistics and ML to compute a score that quantifies the amount of **bias present in a tabular dataset**, furthermore providing evidence that both the accuracy and the unbiasedness of ML models benefit from (proper) data pre-processing.
Finally, in the third study (i.e. the submitted journal paper [9]) we propose a novel technique that combines spectral clustering, an unsupervised ML algorithm, with landmark survival analysis to **identify and characterize subgroups of individuals that differ in terms of survival behaviors** within tabular datasets that include both time-dependent and time-independent features.

1.3 Scientific Collaboration

The research presented in this PhD Thesis was carried on in collaboration with the pneumology unit of the University Hospital of Cattinara, Trieste, which is part of *ASUGI* (Azienda Sanitaria Universitaria Giulioano-Isontina), the Public Health Authority for the provinces of Gorizia and Trieste.

1.4 Funding

This PhD work has been funded by Cassa Risparmio di Trieste, within the Project "Dipartimenti di Eccellenza 2018-2022", project title: "Support to the personalized clinical activity in respiratory medicine via artificial intelligence and machine learning methods."

1.5 Structure of the thesis

The remainder of the thesis is structured as follows:
Chapter 2 provides an overview of the challenges of clinical data, a brief presentation of our COVID-19 dataset, and a panoramic of ML/AI in medicine, focused in particular on topics relevant to this work.
Chapters 3, 4 and 5 present the aforementioned studies [7, 8, 9], describing their methodology and results.
Finally, in Chapter 6, we conclude by summarizing the key findings of our studies, discussing their limitations, and outlining directions for future research.

2 Background

2.1 Challenges of Clinical Data

Clinical data are afflicted by several challenges, due to their sensitive nature, and to the variety of methods, actors and circumstances involved in their life-cycle.

To begin with, collection of clinical data is time consuming and prone to error, as the vast majority of it has to be carried on manually by healthcare workers, often across several meetings with the patients [4, 5, 16]. A first broad distinction can be made between structured and unstructured clinical data. **Structured data** consist of ordered sets of values describing the patient's condition, such as lab results (e.g., levels of C-reactive proteins), manually surveyed information (e.g., age, applied therapies), or medical signals (e.g., air flow/volume/pressure signals recorded with mechanical ventilators, medical imaging), and one of their most common forms is tabular data. **Unstructured data**, instead, comprise free-form text annotated by healthcare workers during conversations with the patients, such as amnestic information collected at hospital admission, or doctors' investigation records. [2]

In modern practice, clinical data are stored onto information systems known as **electronic health records (EHRs)** [1]. The ambitious purpose of these storage systems is providing a unified interface for the acquisition and consultation of patient data, including health as well as administrative information, thus answering the needs of health care practitioners, administrators, patients, researchers, and public monitors [1] . In practice, though, the strict safety requirements, combined with the lack of implementation

[1] In fact, initially EHRs were mainly designed to serve billing purposes, and only later evolved to include clinical details useful for actual healthcare delivery [1].

standards for these systems, often introduce complications during the dataset gathering phase: security protocols, rightfully put in place to safeguard patient's privacy, may not allow exporting data in a tabular format, furthermore requiring several authentications even to access data regarding a single patient, which is often spread across multiple EHRs. Gathering the collected data into datasets for AI and ML development can thus be very challenging, and must often be painstakingly carried on by hand, introducing further occasions for human error to corrupt the data. During this phase, **data missingness** also naturally arises from the fact that patient experiences may vary consistently (especially in multi-centric studies), resulting in tabular datasets with features that cannot be evaluated for all patients.

The **heterogeneity** of clinical information readily translates into datasets that have high dimensionality, and present a variety of data-types [2], typically falling into one of the following classes:

- Numerical - for quantitative data (e.g., age, physiological indicators, days of therapy)
- Categorical - for qualitative data, may present ordered levels (e.g., sex at birth, risk group)
- Binary - indicate if patient meets a given condition (e.g., hypertension) or underwent a specific therapy (e.g., mechanical ventilation)
- Date-time - temporal information (e.g., hospital admission/discharge, start/end of therapy, sampling date)
- Text - unstructured data, or other unique values (e.g., patient ID, pathogen name).

The differences in format and information content among these data-types forces the adoption of elaborate pre-processing strategies, and of models capable of handling this diversity.

Data scarcity and **unbalancedness** of the features' distributions are also common issues in clinical datasets [6, 2], as they naturally arise due to the rarity of some conditions (e.g., rare diseases, low lethality), or due to biases that are inherently present in the sampled population (e.g., older age, presence of minority groups). These are major problems that should be taken into consideration since the design phase of a study, as they negatively impact the performance and generalization capabilities of trained models, and can only be partially solved by adopting resampling strategies.

Finally, the technologies employed in the creation of clinical datasets, as well as the populations and diseases described in them, naturally evolve in time, introducing the so-called **data drifting** and **concept drifting** problems [6]. These are responsible for the obsolescence of trained models, as their ability to make predictions on data that are more recent than the training data may deteriorate, or be completely shut down, thus requiring to periodically update or completely retrain the models.

These challenges provide a strong motivation for the development of AI/ML techniques that are versatile, understandable, and effective, as well as of tools to evaluate the quality of clinical datasets employed for this purpose.

2.2 COVID-19 dataset

The raw dataset comprises data regarding **951 COVID-19 patients** which were hospitalized across 26 Italian centers between February 2020 and May 2023. It was extracted from a pool of 1012 patients, obtained by combining the datasets of two former randomized trial studies investigating the efficacy of steroid treatments on patients affected by SARS-COV-2-induced pneumonia [17, 18]. Data gathering and part of the data collection were performed by our collaborators from the pneumology department of the university hospital of Cattinara, in Trieste, who also authored these previous studies.

The raw dataset comprises the following 129 features:

- **health history** (24) - features collected at hospital admission, these include age, sex, body-mass index, and 21 risk factors, namely, smoking habits, COVID-19 vaccination status, and past occurrence of the following conditions: diabetes, oncologic diseases, hypertension, asthma, chronic obstructive pulmonary disease (COPD), bronchiectasis, dyslipidemia, chronic renal disease (CRD), atrial fibrillation, coronary artery disease (CAD), chronic heart failure, cardiovascular diseases, obstructive sleep apnea, cerebrovascular diseases, dementia, polycythemia, latent tuberculosis, autoimmune diseases, and pulmonary embolism. All of these features except the first four mentioned are of binary datatype.

- **therapy** (20) - binary features detailing which therapies were administered to the patient, including ventilatory supports (high flow nasal cannula (HFNC), non invasive ventilation (NIV), and invasive mechanical ventilation (IMV)), clinical procedures (pronation, tracheostomy, and extra-corporeal membrane oxygenation (ECMO) and medications (warfarin, new oral anticoagulants, tocilizumab, remdesvir, monoclonal antibodies, heparine (non fractioned or low molecular weight with prophylactic or anticoagulant dosages), and steroid therapy (with additional features specifying the treatment group, and count of days of full dosage)).

- **complications** (21) - binary features detailing which complications occurred during the treatment of the patients. These include: bradycardia, diarrhea, increased liver enzymes, hypotension, hypokalemia, hypernatremia, shock, acute renal failure (ARF), intravascular coagulation, acute heart failure, stroke, atrial fibrillation, superinfection (and annex text feature with the micro-organism specification), hyperglycemia, respiratory acidosis, pneumothorax (PNX), pneumomediastium, critical illness myopathy, pulmonary embolism, and cardio-respiratory arrest.

- **serial measurements** (45) - features comprising (up to) five measurements for each

of nine variables, including: type of ventilatory support, sequential organ failure assessment score (SOFA), and levels of arterial partial pressure of oxygen to fraction of inspired oxygen ratio (PaO_2/FiO_2), C-reactive Protein (CRP), lymphocytes, lactase, lactate dehydrogenase (LDH), and D-dimer.

- **dates** (18) - these include dates of birth, hospitalization, decease/discharge, enrollment, symptoms onset, sampling of serial values, start/end of NIV, IMV and steroid therapy;
- **outcome** (1) - binary indicator of patient's death.

Of these, only **79 features** could effectively be considered for data analysis, as others were either excessively sparse (e.g., 7 out of the 9 serially measured features), or presented uninformative distributions (e.g. same value for $> 95\%$ of the samples). The dataset is **heterogeneous**, as it includes features of all the data-types mentioned in section 2.1, with the majority of them being binary. Time information, present in the data in the form of dates, necessitated extensive manual intervention to identify and correct inconsistencies, furthermore requiring the introduction of specific strategies to be coupled with the pertinent features, and integrated into the models. Serial measurements, in particular, vary both in quantity (from 0 to 5 samples) and exact timing across the patients. Furthermore, the dataset is **unbalanced** with respect to the outcome, as 16% of the patients in the dataset died. Details on the distributions of the features are discussed within the contexts of the specific studies, presented in the following chapters.

2.3 Machine Learning in Medicine

Machine learning (ML) algorithms are computer programs that acquire the ability to perform tasks by training onto data, and constitute the backbone of modern artificial intelligence (AI). In the medical domain, ML enables the automatic extraction of patterns from complex, multi-modal datasets, supporting tasks that range from diagnosis and prognosis to treatment recommendation and resource allocation. It is possible to distinguish different families of ML approaches, basing on the tasks they can tackle, and on the data required for training them. For the purpose of this work, we will briefly discuss supervised and unsupervised approaches.

2.3.1 Supervised Learning

Supervised learning approaches extrapolate input-output mappings present in the training data, producing models that are capable of predicting outputs for previously unseen inputs. These algorithms are employed in prediction tasks, whose main instances are:

- **Classification**, when the outputs are of binary or categorical data-type, e.g., to predict if a patient with given features will survive or not, as in our study detailed in Chapter 3

- **Regression**, when the outputs are quantitative, e.g., to estimate the level of a given biomarker in patient's blood.

A large variety of algorithms belong to this family of approaches, including linear models (e.g., logistic regression), tree-based ensembles (random forests, gradient boosting), kernel methods (support vector machines), and neural networks. Supervised learning methods have been widely employed in medical research and clinical applications. Support Vector Machines (SVMs) have been used for diagnosing pulmonary hypertension [19] and predicting optimal outcomes in intensive care settings [20], while Random Forests (RFs) have shown effectiveness in analyzing gene expression data [21]. Convolutional Neural Networks (CNNs) have also achieved remarkable success in medical imaging and signal analysis, enabling the development of AI tools for screening breast cancer in mammograms [22], diagnosing diabetic retinopathy from fundus photographs [23], classifying prostate cancer using histopathologic whole-slide images [24], and detecting cardiac arrhythmias from electrocardiogram signals [25].

Because of the challenges presented by clinical data (discussed in Section 2.1), careful **validation** of the models' generalization capabilities is essential [26]. Cross-validation, stratified sampling, and bootstrap resampling are commonly employed to ensure that metrics such as accuracy, sensitivity, specificity, Area Under the Receiver Operating Characteristic (AUROC), and Area Under the Precision–Recall Curve (AUPRC) provide unbiased estimates of the models' performance outside the training set [27].

2.3.2 Unsupervised Learning

In contrast, unsupervised learning approaches do not rely on labels in the training data. Instead, they exploit complex patterns and relations among the features, producing models that compute an alternative representation of the data, thus revealing the underlying structure of the dataset. The typical use of these approaches concerns inferential and exploratory tasks, such as:

- **Clustering**, which consists in finding meaningful partitions of the input data, e.g., to classify patients into risk groups, as in our study detailed in Chapter 5;
- **Anomaly detection**, which identifies outliers or rare cases, as done by our tool presented in Chapter 4;
- **Dimensionality Reduction**, that is, computing low-dimensional projections of high-dimensional data while preserving their essential characteristics, often to facilitate visualization, or to improve downstream performance in supervised tasks.

Common unsupervised learning algorithms include spectral clustering, hierarchical clustering, and k-means for clustering, principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) for dimensionality reduction, and isolation forests and autoencoders for

anomaly detection algorithms. These methods have been applied across diverse biomedical domains, including clustering analysis of imaging phenotypes [28], exploration of gene expression profiles [29], discovery of novel features, protein signatures, and biomarkers for disease diagnosis and monitoring [30], and identification of previously unrecognized disease subtypes or patient subpopulations that may benefit from personalized treatment strategies [31].

2.3.3 Explainable AI

A key consideration in applying ML to healthcare is interpretability. Clinicians must understand the reasoning behind model outputs, especially when results could influence patient treatment. **Explainable Artificial Intelligence (XAI)** addresses this need by providing tools that make model outputs comprehensible to human experts, enabling clinicians to interpret predictions in relation to their own knowledge and experience [32]. Among the most widely adopted approaches, Shapley Additive Explanations (SHAP) [33] offer a unified framework for quantifying the contribution of each feature to individual predictions. By decomposing a model's output into additive components, SHAP allows practitioners to verify whether the model relies on clinically meaningful variables, exposing potential biases or inconsistencies in its reasoning. These explanations enhance trust in model outputs, facilitate validation, and support the integration of data-driven systems into clinical workflows. Explainable AI thus serves as a bridge between computational inference and medical reasoning, ensuring that predictive models not only achieve high performance but also provide insights that are interpretable, verifiable, and aligned with clinical understanding.

2.3.4 Generative AI/ML

Generative ML represents a rapidly developing branch of AI aimed at producing new data that resemble existing distributions, such as text, images, or multi-modal content [34, 35]. Recent advances—particularly in transformer-based large language models (LLMs), generative adversarial networks (GANs), and diffusion models have greatly expanded their applicability in medicine [36]. LLMs can summarize clinical documentation, generate standardized diagnostic reports, and facilitate medical education through interactive dialogue systems [37]. Similarly, image-based generative models can create synthetic pathology or radiology images, enriching limited datasets and enhancing model robustness [38]. Moreover, multi-modal and multi-agent frameworks have been employed to integrate textual, imaging, and molecular data into unified diagnostic and decision-support tools [39].

Despite their promising capabilities, generative models raise concerns related to data validity, bias propagation, and interpretability, which remain active research challenges [40, 41]. Furthermore, the development and deployment of such models require massive datasets and substantial computational resources, positioning them within a research venue that lies beyond the scope of the present work.

2.3.5 Further Reading

Details on the specific techniques used in this work are discussed within the context of the pertinent studies, presented in the following chapters. For an in-depth discussion on the technical aspects of ML, the reader is addressed to the excellent manuals edited by Kevin Murphy [26]; furthermore, the series of papers edited by Rashidi et al. [42] provides a comprehensive review of the current state of ML and AI in the clinical context.

3

A tailored machine learning approach for mortality prediction in severe COVID-19 treated with glucocorticoids.

The present chapter is based on the paper "A tailored machine learning approach for mortality prediction in severe COVID-19 treated with glucocorticoids", published on The International Journal of Tuberculosis and Lung Disease 28.9 (2024): 439-445, of which I am co-first author [7].

3.1 Introduction

Severe pneumonia in COVID-19 patients poses a unique challenge to healthcare systems, presenting a critical gap in the identification of individuals who undergo rapid deterioration, leading to unfavorable outcomes and post-acute sequelae [43]. Glucocorticoids (GCs) are the most studied and effective agents for severe SARS-CoV-2-related pneumonia [44]. However, a current challenge of using GCs in severe COVID-19 arises from the inability to predict non-responsive patients, hindering the timely escalation of respiratory support and personalized pharmacological interventions. Indeed, it has been suggested that adjusting the dose and duration of GC treatment according to clinical progression may ameliorate outcomes [18]. While clinical prediction scores exist for risk stratification in community-acquired pneumonia and COVID-19, they often rely on predefined clinical parameters that may not fully capture the dynamic nature of pneumonia progression, nor are they designed to predict the efficacy of pharmacological treatments like GCs [45, 46]. The recent

diffusion of artificial intelligence (AI), especially machine learning (ML) techniques, has enabled models to exhibit unprecedented predictive capabilities and adaptability to different data types [47]. These advancements represent a stride towards providing healthcare professionals with a useful tool to assist in clinical decision-making at the bedside, which is crucial for obtaining more accurate forecasts regarding outcomes linked to the usage of specific drugs in various clinical scenarios. While predictive ML solutions have been proposed in clinical studies, to the best of our knowledge, no association between anamnestic or therapeutic features, and adverse outcomes in a population of severe COVID-19 patients treated with GCs has been explored [48]. Starting from a large dataset of patients with COVID-19 pneumonia, we implemented an ML pipeline to train and validate models capable of integrating baseline clinical and laboratory data. Our aim was to effectively predict the risk of death of patients treated with GCs and quantitatively estimate the impact of each feature on individual predictions.

3.2 Methods

3.2.1 Data

Starting from the raw dataset, presented in section 2.2, we included only the patients who underwent GC treatment for which outcome was known, obtaining a pool of **825 patients**. We analyzed **52 features**, including demographic information and data related to medical history, therapy, and complications. Serial measurements were also available for the arterial partial pressure of oxygen to fraction of inspired oxygen ratio (PaO_2/FiO_2, mmHg) and C-reactive protein levels (CRP, mg/L) at Days 0, 3, 7, 14 and 28 from hospitalization. We employed the earliest available measurement to deal with these time series and applied expert-designed heuristics to determine if the data improved over time. We classified a time series as 'improving' in the following cases:

1. for PaO_2/FiO_2, if the latest value was above 250 mmHg or if it increased by at least 40% compared to the earliest value;
2. for CRP level, if the latest value was below 10 mg/L, or if it decreased by at least 40% compared to the earliest value.

This study was approved by both the Italian National Ethics Committee, Rome, Italy (approval number 2020-006054-43, 2 January 2021) and the referral local Ethics Committee (approval number CEUR-2020-Os-052, 23 March 2020).

3.2.2 Machine Learning Methods

We developed a pipeline to train different ML algorithms to predict patient death during hospitalization and to evaluate them (figure 3.1). The task is an instance of a supervised binary classification problem on tabular data. We trained and evaluated six ML algorithms from the most commonly employed ones in clinical medicine, i.e. logistic regression (LR),

Variable	Distribution Statistic	
	n (%)	Missing Values n(%)
General information		
Patients, n	825	
Days hospitalized, median [IQR]	16 [11–23]	9 (1.1)
Days to enrollment, median [IQR]	1 [0–1]	65 (7.9)
Outcome: hospital death	118 (14.3)	0
Medical history		
Age, years, median [IQR]	65 [55–74]	0
Sex		0
Male	574 (69.6)	
Female	251 (30.4)	
Body mass index, kg/m2, median [IQR]	27.4 [24.7–30.7]	123 (14.9)
Coronary artery disease	72 (8.7)	8 (1.0)
Chronic heart failure	64 (7.8)	11 (1.3)
Therapy and complications		
Invasive mechanical ventilation	113 (13.7)	1 (0.1)
Days of steroid therapy, median [IQR]	10 [8–11]	124 (15.0)
Acute renal failure	54 (6.6)	53 (6.4)
Extracted from time series		
PaO_2/FiO_2 improving	447 (54.2)	0
CRP improving	585 (70.9)	0
PaO_2/FiO_2 earliest read, median [IQR]	180.0 [122.65–251.68]	33 (4)
CRP earliest read, median [IQR]	77.95 [37.0–128.1]	33 (4)

Table 3.1: Final dataset, including details about the distribution of the main features (as determined during the feature selection step of the pipeline).

support vector machine (SVM), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost, XGB) and a fully-connected feed-forward neural network, also known as multi-layer perceptron (MLP). The following paragraphs provide a brief review of these algorithms; for a more in-depth discussion, the reader is addressed to [49, 26].

3.2.2.1 Logistic regression (LR)

This is the classic machine learning algorithm for binary classification. It belongs to the wider class of generalized linear models (GLMs), as the model it fits computes a linear combination of the components of the input, and applies a non-linear function to it, in this case a logistic function, returning a value in the interval $[0, 1]$ that can be interpreted as a probability. The model is trained using maximum likelihood estimation, which consists in finding the values of the parameters (i.e. the weights) which maximize the likelihood of observing the training data. The main hyper-parameters of the model are the choice of the solving algorithm, which determines how the optimal parameters are actually computed, and the choice of the regularization technique, which in this case consists in including an extra term in the computation of the likelihood that penalizes the choice of larger weights, effectively reducing overfitting to the training data, and thus improving the model's ability to generalize. Although easy to interpret, the resulting model is not able to capture more complex distributions.

3.2.2.2 Decision tree (DT)

As the name suggests, these models have a tree-like structure in which nodes correspond to subsequent splits of the input space; input vectors are mapped to outputs by iteratively comparing the values of their features against the splitting conditions associated with each node of the tree, starting from its root and until they reach a leaf node, which is associated with a specific output value (in our case, either 0 or 1). The training process consists in iteratively splitting the training data by maximizing the separation of the resulting splits with respect to the output classes; this is achieved using a greedy approach, that is, choosing the single best variable to define the splitting condition at each step. The resulting model is very easy to understand, since any output corresponds to a sequence of answers to yes/no questions about the values of the input, and it is capable of fitting even very complex distributions. Unfortunately these models also have a tendency to overfit the training data; the maximum depth of the tree and the minimum number of samples per leaf can be fine tuned to reduce this effect and improve the generalization capabilities of the model.

3.2.2.3 Random forest (RF)

Random forests are ensemble models constituted by a collection of decision trees. Their predictions are obtained by aggregating the outputs of several decision trees evaluated on the same input [50]. Each of these trees is trained on a bootstrap dataset, obtained by randomly resampling the original dataset with reinsertion, and keeping only a random subset of the input features; due to limited access to the training data, the accuracy of the individual trees is low, though aggregation has the effect of balancing their errors, resulting in a model which is more accurate and less prone to overfit than a single decision tree. The model can be tuned by changing the total number of trees in the forest, the numbers of rows and features that will appear in each of the bootstrapped datasets, and the parameters for fitting each of the trees.

3.2.2.4 Extreme Gradient Boosting (XGBoost)

XGBoost is yet another ensemble model based on decision trees [51]. Its predictions are obtained by computing a linear combination of the of the outputs of several decision trees; unlike random forests, in which all the trees are trained to perform the same task, in this case each tree is trained to produce the best adjustment to the ensemble prediction of its predecessors (= boosting), whereas the adjustments themselves are recomputed at each step by optimizing a measure of the training error through gradient optimization methods (e.g., SGD, adam); The weight of the contribution of each tree in the final prediction is determined both by a prefixed learning rate and the performance of the individual tree on a holdout set, which is estimated during training (more specifically, while performing gradient optimization, as it is necessary for early stopping). Several parameters can be adjusted to tune these models: beside those that we already mentioned for the single

trees, it is possible to tweak the learning rate, choose different regularization term to add to the loss, and pick and adjust the chosen SGD variant, which comes with its own set of parameters.

3.2.2.5 Support Vector Machines (SVM)

Support vector machines [52] compute an optimal boundary between regions of the input spaces belonging to distinct classes. Predictions are based on the distance of the input vector from this surface, which is estimated by computing a weighted sum of some terms that represent a sort of distance between the input vector and a set of so-called "support vectors", which concretely define the boundary. Model fitting thus consists in identifying these support vectors within the training data and choosing their weight for the final computation. The process involves the use of a kernel function, which maps the vectors into a higher dimension space in which the desired boundary is a hyperplane. The weights, and thus the hyperplane, are determined by optimizing a loss function which depends on the size of the margin, that is, the boundary region. The hyper-parameters of the model include the choice of the kernel function and the associated parameters, regularization term/parameters, and optimization algorithm (e.g., SVG).

3.2.2.6 Multilayer Perceptron (MLP)

These models belong to the family of neural networks, that is, models whose mathematical formulation is more easily described with the use of a directed graph; that of an MLP is structured as a series of subsequent layers, in which every node is connected to all the nodes in the subsequent layer, representing the fact the the input vector is iteratively processed trough a series of parallel computations, resulting in an output vector of the desired shape (in our case, a single number in the interval [0, 1]). Each node computes a linear combination of its inputs and feeds it into a non-linear activation function, which isn't different from what happens in logistic regression. The computation performed by the whole network though is much harder to comprehend intuitively, as several of these simpler computations are arranged in a nested fashion, involving a large amount of parameters. Although scarcely interpretable, these models are highly versatile and able to approximate any distribution (within the limits of the available data and computational resources). The values of the parameters (i.e. the weights and biases of the linear combinations) are computed by optimizing a loss function which measures the error in fitting the training data (e.g., Mean Squared Error, Cross Entropy), typically using variations of SGD. Depending on the complexity of the problem and of the chosen network architecture, this process can get quite intense in terms of computational resources. Fine-tuning these models can thus be very challenging, both because of the ample variety of choices available in terms of hyper-parameters (number and size of layers, activation functions, optimization algorithm, loss function, regularization), and because of the non-negligible costs required to evaluate the performance of any chosen configuration.

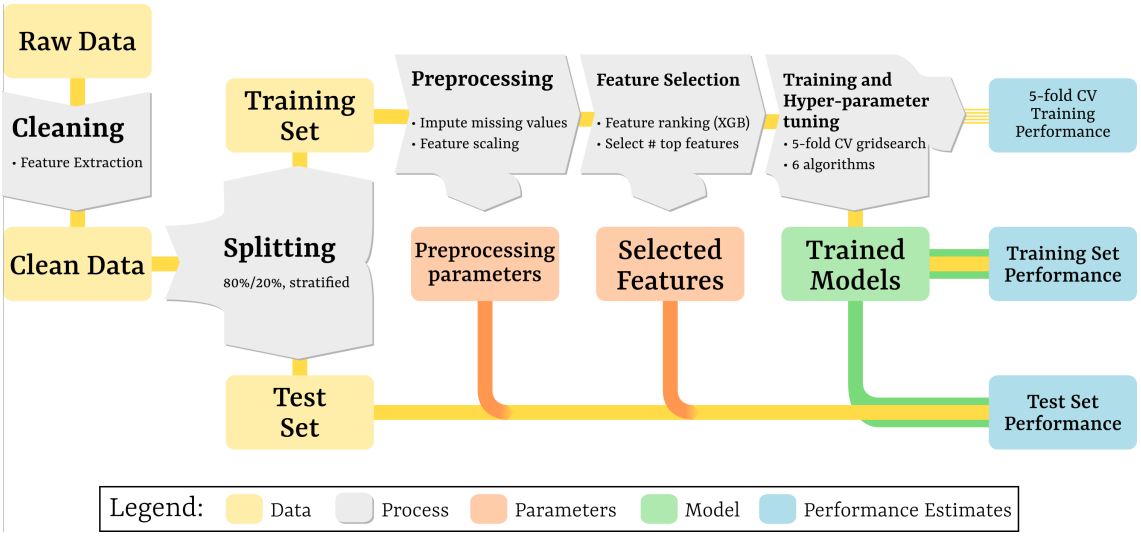

Figure 3.1: Flow chart showing the machine learning pipeline developed to train all machine learning algorithms and to evaluate their performance.

3.2.3 Training set and Validation

The dataset is randomly split into training (80%) and test (20%) sets, enforcing stratification to preserve the outcome proportions. The test set was not used in any step except internal validation to obtain an unbiased estimate of the final models' performance.

3.2.4 Evaluation Metrics

We use the areas under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) to measure the performance of our models. AUPRC is computed similarly to AUROC, replacing the false-positive rate with precision (i.e. positive predictive value). While less popular than the latter, it has been regarded as a better choice for tasks on imbalanced datasets [27]. For this reason, we used AUPRC to determine the best models during variable selection and hyperparameter tuning. To evaluate the final models, the AUROC and AUPRC were computed on both the training set and the test set, with bootstrap sampling (with 1,000 samples) adopted to provide a 95% confidence interval of the considered metrics.

3.2.5 Preprocessing

Preprocessing consists of two steps: missing value imputation and feature scaling [53]. In the former step, we compute the mean of numeric features and the mode of binary/categorical features over the training set and used these to fill in the missing values. We then apply feature scaling by dividing the real-valued features by their standard deviation over the training set, a common procedure for ML algorithms such as SVM and NN.

3.2.6 Feature Selection

The feature selection step allows us to determine which variables are most important for the classification task. In line with the literature, we perform feature selection by ranking them and selecting the best ones to be included in the final models [54]. We employ XGB estimates of feature importance to rank the features, which are computed during model fitting by evaluating how often each feature contributes to correct predictions. We fit an XGB model (hyper-parameters: maximum depth=3, learning rate=0.2, and alpha=5) on 100 random sub-samples, each consisting of 80% of the training set. We then average the estimated importances and rank the features accordingly. After ranking the features, the selection of the optimal number of features also relies on XGB: this involves fitting the model, with the same hyper-parameters, on the training set while incrementally adding features from one to 20 in the sequence provided by the ranking, and employing a five-fold cross-validation (CV) scheme to estimate the associated generalization metrics [49]. XGB was chosen for these tasks as both the literature and preliminary tests suggested its superiority compared to simpler neural network architectures for our case [55]. The final feature sets are selected to maximize the estimated AUPRC and AUROC while keeping the number of features as low as possible, thus reducing the risk of overfitting and improving the interpretability of the models, as well as training costs and overall performance [26].

3.2.7 Training

The best hyper-parameters for each algorithm are determined by grid search [26]. A single round consisted of training several configurations of the same algorithm, each corresponding to a point in the hyper-parameter space, and estimating their generalization performance on the training set, employing a five-fold cross-validation scheme. Each algorithm's best scoring configuration across all rounds is picked. .

3.2.8 Model Explanation

We compute the Shapley additive explanation values (SHAP) for the predictions of the best models on the training data [33]. SHAP values represent a powerful and versatile tool that can be applied to explain how a model predicts the outcome for any given sample. Each feature of each sample (i.e. patient) is associated with its own SHAP value, estimating how that specific value affects the prediction of the model. The sum of the SHAP values for a given sample corresponds to the difference between the prediction for that sample and the average prediction of the model. In other words, SHAP values quantify how much—and in which direction—each feature contributes to pushing the model's prediction from its "default guess" towards its final value. In our case, a positive SHAP value indicates that the corresponding feature value increases the probability of death during hospitalization.

Model	Training metrics		Test metrics	
	AUROC	AUPRC	AUROC	AUPRC
LR	0.905 (0.882–0.926)	0.632 (0.553–0.709)	0.818 (0.7–0.924)	0.645 (0.467–0.792)
SVM	0.905 (0.881–0.927)	0.631 (0.549–0.709)	0.882 (0.812–0.94)	0.651 (0.458–0.81)
DT	0.851 (0.825–0.876)	0.632 (0.584–0.679)	0.888 (0.847–0.927)	0.69 (0.596–0.772
RF	0.934 (0.917–0.949)	0.704 (0.627–0.774)	0.938 (0.903–0.969)	0.714 (0.548–0.856)
XGB	0.957 (0.942–0.97)	0.776 (0.699–0.848)	0.937 (0.901–0.968)	0.701 (0.538–0.846)
NN	0.915 (0.896–0.933)	0.665 (0.601–0.724)	0.875 (0.829–0.916)	0.652 (0.549–0.745)

Table 3.2: Training and test metrics of the final models (Mean (CI)). Models in their final configurations are trained on the whole training set, metrics are evaluated on 1,000 bootstrap samples of the training and test sets, respectively.
LR=Logistic Regression; SVM=Support vector machine; DT=Decision Tree; RF= Random Forest; XGB=Extreme Gradient Boosting; NN=Neural Network; AUROC=area under the receiver operating characteristic curve; AUPRC=area under the precision-recall curve.

3.2.9 Implementation Details

All the code to perform the analysis was written in Python v3.10.8. We employed the ML algorithm implementation provided in scikit-learn v1.2.0 (LR, SVM, DT, RF, NN) [56] and XGBoost v1.7.3,[57] while SHAP v0.42.1 was used to compute SHAP values [58].

3.3 Results

We identified **9 features** from the available 52 through the previously described selection strategy. Following their ranking order, these were: CRP improvement, PaO_2/FiO_2 improvement, age, coronary artery disease, need for invasive mechanical ventilation (IMV), acute renal failure, chronic heart failure, PaO_2/FiO_2 ratio earliest value, and body mass index (BMI). Table 3.1 describes these features and the main characteristics of the study population.

Random forest (RF) achieved the highest performance, scoring an AUROC of 0.938 (95%confidence interval [CI] 0.903–0.969) on the test set. XGB followed with a test AUROC of 0.937 (95% CI 0.901–0.968). DT obtained the third-best test AUROC (0.888, 95% CI 0.847–0.927), suggesting that tree-based methods were the most appropriate in our setting. 3.2 shows the training and test metrics of all investigated models.

The SHAP values for the best model (i.e. RF) indicated that PaO_2/FiO_2 ratio improvement and age had the most substantial impact on the final predictions (see Figure 3.2) , their mean absolute SHAP being 0.149 and 0.13, respectively . Indeed, the model found a strong positive correlation between age and outcome and a strong inverse correlation between PaO_2/FiO_2 ratio improvement (average RF SHAP of –0.19 for improving patients, +0.1 for the others) and outcome. CRP improvement and the earliest available PaO_2/FiO_2 value were the most impactful features (mean absolute RF SHAP 0.071 and 0.045), presenting a negative correlation with the outcome. The use of IMV also positively correlated with the outcome, although with a progressively decreasing magnitude

of impact (0.027) and a more pronounced harmful effect (+0.07 average RF SHAP for patients undergoing IMV vs –0.02 for those who did not require IMV). Concerning BMI, the model associated values between 24 and 29 with a slight decrease in the probability of death (–0.01 average RF SHAP), extremely low values with a slight increase (up to +0.02 average RF SHAP), and extremely high values with an apparently protective effect. Notably, XGB (i.e. the second-best model) only disagreed with RF on this last association , most likely due to the scarcity of patients with extremely high BMIs. Altogether, BMI and the remaining features (i.e. acute renal failure, coronary artery disease and chronic heart failure) presented a relatively small average absolute impact (<0.009).

3.4 Discussion and Conclusions

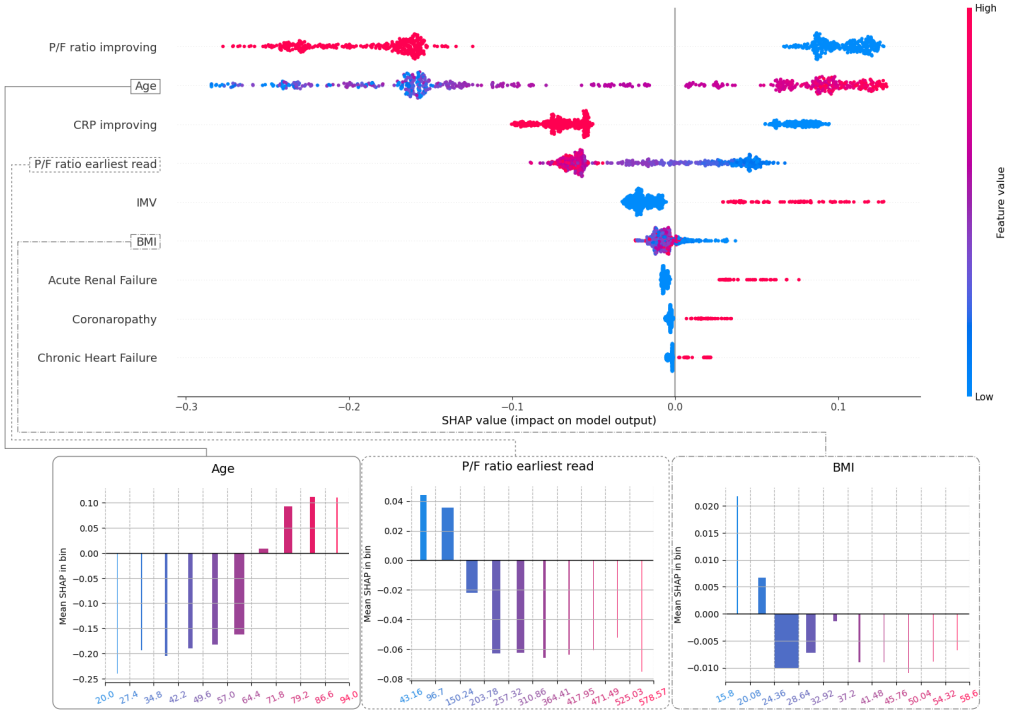

Figure 3.2: SHAP values for random forest on the training set. In the upper panel, the SHAP values for each feature of each patient are displayed as dots on separate stacked horizontal lines, with the color representing the value of the features (same as in the bottom panels). SHAP values are expressed as probability variations in $[-1, 1]$ and features are ordered according to their average absolute SHAP, which indicates the magnitude of the overall effect of a feature on the model's predictions. In the bottom panels, the average SHAP values of the numeric features (age, PaO_2/FiO_2 ratio earliest value, and BMI) are reported over 10 equispaced bins. Bar heights (y axis) represent the average SHAP value in the bin; bar widths scale linearly with the number of patients in the bin (minimum 1, maximum 339).
CRP=C-reactive protein; IMV=invasive mechanical ventilation; BMI=body mass index; SHAP=Shapley additive explanation; PaO_2/FiO_2=arterial partial pressure of oxygen to fraction of inspired oxygen ratio.

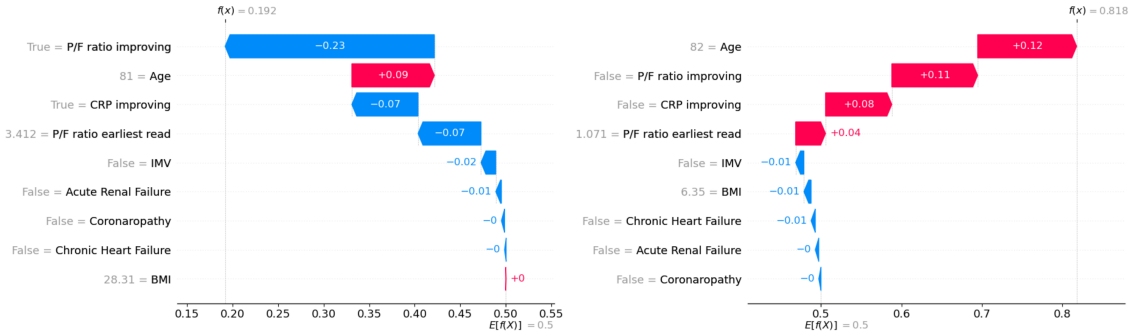

Figure 3.3: Waterfall plots showing the SHAP values for all the features of two sample patients offer a local explanation of the model's output, estimating the impact of each feature on the outcome predicted by the algorithm (random forest). The color represents the sign of the SHAP contribution. The SHAP values attributed to the single features of the patients are sorted by the most influential factors for individual prediction.
Left panel: Prediction for a survived patient; Right panel: prediction for a dead patient. CRP=C-reactive protein; IMV=invasive mechanical ventilation; BMI=body mass index; SHAP=Shapley additive explanation.

We employed ML techniques to predict hospital mortality among individuals experiencing SARS-CoV-2-related pneumonia treated with GCs per the WHO recommendations. Unlike classical statistical analysis, which posits assumptions on the shape of the relationship between variables and outcome, ML models can learn complex (i.e. highly non-linear) relations in the data, with a development process that prioritizes the quality of prediction. Our approach involved an initial ML algorithm (XGB) to identify the most relevant features for mortality prediction from a pool of 52 variables. The available data included four features derived from time-course data: notably, three of these (i.e. CRP improvement, PaO_2/FiO_2 improvement and PaO_2/FiO_2 earliest read) were found particularly effective for the task. In contrast, the earliest CRP read was excluded. We fitted and procedurally optimized the hyper-parameters of six ML algorithms belonging to different classes to determine which was most appropriate for our data. RF emerged with the highest test AUROC, closely followed by XGB. The performance of each algorithm underwent blind internal validation on a test cohort of patients. We then computed SHAP values for both RF and XGB, obtaining a detailed account of each prediction. These values not only allow us to spot relations (not limited to linear correlation) between a given feature and the outcome globally (i.e. across the whole training dataset) but also to estimate the effect of features locally (i.e. at a single prediction level), which improves the interpretability of the model's behavior. Indeed, SHAP values offer more insightful information than p-values derived from traditional statistical testing, as the latter provide a singular numerical estimate of statistical confidence in hypotheses related to rather generic aspects of the relationship between variables and outcomes. We found that improving PaO_2/FiO_2 and CRP readings, younger age, absence of comorbidities and the need for IMV were predictive factors for a higher likelihood of survival. Notably, the earliest PaO_2/FiO_2 reading demonstrated less impact than PaO_2/FiO_2 improvement over time, suggesting the need for close monitoring

and prompt escalation of respiratory support if the trend shows no improvement or deterioration. These results also indicate that the improvement of PaO_2/FiO_2 is more relevant than the improvement of CRP in determining the prognosis of patients treated with prolonged GCs, which has not been previously reported. Indeed, a strength of ML techniques is their ability to surpass prevailing clinical assumptions [59]. Limited prior research has assessed the use of ML methodologies to develop predictive models for patients with severe COVID-19 pneumonia [60, 61, 62, 63, 64]. In contrast to our report, these studies often achieved lower performance levels, did not consistently focus on patients with respiratory involvement due to COVID-19 pneumonia and did not selectively include patients treated with GCs [65]. Additionally, all the algorithms were exclusively trained on baseline data, neglecting the incorporation of time-course data crucial in critical care scenarios.

Our study exhibits some limitations. As in many ML-driven analyses, understanding each step of the algorithm's decision-making process is a challenge that is not entirely overcome using SHAP values, which explain the final prediction but not all the internal steps. However, this can also be considered a necessary trade-off since some complex relations can only be handled with algorithms producing non-fully explainable models. Furthermore, despite our study centered on a population undergoing GCs, our results cannot differentiate between features predicting GC treatment failure and those contributing to an unfavorable outcome due to factors independent from GC susceptibility. As a final remark, our algorithm does not provide specific recommendations for treatment adjustments, even if previous literature demonstrated that the timely escalation of dose and duration of GC treatment in patients who deteriorate may enhance survival [66]. However, this dynamic model can adjust its predictions based on time-course data inputted during clinical re-evaluations. This allows clinicians to assess the potential benefits of adopted management changes over time. Our ML tool is not intended to relieve clinicians of their responsibilities but to empower them in their daily practice to increase awareness and attentiveness to the potential consequences of their therapeutic choices. In conclusion, the algorithm we trained obtained a high AUROC and underwent blind validation, paving the way to its future practical and customizable application in real-world clinical settings. Our ML pipeline can be adapted to datasets with similar clinical features, ensuring the extraction of reliable insights that align with the unique characteristics of each study cohort.

4

Investigating Fairness with FanFAIR: is Pre-processing Useful Only for Performances?

The present chapter is based on the conference paper "Investigating Fairness with FanFAIR: is Pre-Processing Useful Only for Performances?", published on the proceedings of 2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM), IEEE, 2025, of which I am first author [8].

4.1 Introduction

As Artificial Intelligence (AI) continues to be increasingly employed across different domains, especially in high-stakes fields such as healthcare, discussion over its societal impact has gradually become more present in the literature. One of the main concerns in AI and Machine Learning (ML) ethics is the notion of fairness concerning datasets used in training AI models [67]. Fairness, in this context, relates to mitigating biases embedded in datasets to prevent AI systems from perpetuating or exacerbating inequities. While technical methods like debiasing have been advanced to address these challenges, a broader debate questions the effectiveness of such approaches, raising fundamental concerns about their ability to address systemic inequalities.

At the core of dataset fairness is the recognition that datasets can encode historical and structural biases that reflect an unfair world [68]. Such biases are shaped by gendered, racial, colonial, and other discriminatory practices that are embedded in healthcare, employment, and other societal structures [69].

The European Union has taken steps to address the issue of fairness in AI systems, such

as enacting provisions that mandate fairness, transparency, and accountability in AI design [70]. However, as the European Digital Rights (EDRi) report highlights, these policies often take a limited view of fairness, focusing on debiasing data without addressing the larger societal context in which AI systems are deployed. Furthermore, the report suggest that even efforts to conform to the AI Act ensuring that the datasets are "representative, error-free, and complete" may still result in AI systems that reflect an inequitable world [68].

This highlights a significant tension: while datasets can be technically debiased – according to some predefined fairness metrics – they may still perpetuate and amplify the injustices inherent in the systems from which they were derived. In line with this critique, Hanna et al. emphasize the importance of reframing discussions about fairness in AI away from the algorithmic level and toward the social and institutional contexts in which these systems are implemented [71]. Nevertheless, both the EDRi report and the literature recognize the importance of data quality and statistical considerations pertaining to data, as those significantly influence the quality of the ML model [72, 73, 74]. This tension between the goals of dataset fairness and the realities of an unjust world points to a need for broader, more comprehensive approaches to fairness in AI. It is not enough to focus solely on the technical aspects of fairness – such as ensuring that datasets are balanced and free from errors – without also considering the social, political, and economic systems that shape AI's development and deployment. Addressing fairness in AI requires engaging with these larger systems of power and inequality, rather than relying solely on technical solutions to solve deeply entrenched social problems.

More concretely, we believe that unfairness is not solely depending on the data, but it extends to several human activities that impact how data is collected and processed. Some examples are ensuring the consent to collect and reuse patients' personal data [75], enacting the transparency principle over the reuse of data [76], and performing an extensive ethics assessment, which is much broader than just obtaining the permission from an Ethical Review Board [77]. The ethics assessment includes abiding to principles such as Accountability, Dignity and Self-Determination, Traceability, Involvement of Stakeholders, Risk Assessment, and Impact on Society, which are important evaluations for the whole AI life cycle and can help in mitigating and preventing several AI harms.

FanFAIR was published in this context: a software solution for the evaluation of (medical) datasets, that exploits a rule-based fuzzy inference system to provide a quick, semi-automatic assessment of the fairness of data [78]. In FanFAIR, all the aforementioned literature was considered by integrating statistical properties of the dataset, which are computed autonomously by our software, with qualitative considerations entered by the user. FanFAIR can be useful to decide how to pre-process the dataset, and even to decide whether discarding the dataset altogether would be the most appropriate choice, in order to avoid the risk of increasing AI harm. We thus believe that the assessment provided by FanFAIR may serve as the basis for the dataset evaluation, offering helpful insights since the earliest stages of development of AI systems.

In this work, we extend FanFAIR with additional functionalities to simplify data import (Section 4.2.2.3) and to perform an improved outliers detection also in presence of missing data (Section 4.2.2.2). We also introduce a novel facility for the analysis of sensitive variables, which was integrated with the fuzzy inference system (Section 4.2.2.1). Additionally, we test our improved method on a real world dataset about COVID patients, presenting a concrete example of analysis performed with FanFAIR, and discussing the impact of pre-processing on fairness and predicting performance of ML models trained on our data (Section 4.2.3). After presenting our results in Section 4.3, we conclude the manuscript with some comments about limitations of FanFAIR, and possible future developments.

Both the source code and the documentation for FanFAIR are available on GITHUB: https://github.com/aresio/FanFAIR. FanFAIR can also be installed using pip.

4.2 Methods

4.2.1 FanFAIR

FanFAIR is a Python library designed for the semi-automatic assessment of dataset fairness using a rule-based approach that leverages fuzzy logic [78]. The tool calculates multiple fairness metrics over a dataset, and combines them into a single score, which enables researchers to evaluate a dataset's fairness more efficiently. The metrics considered by FanFAIR for the assessment of a dataset are the following [78]:

- *balance*: how balanced the dataset is, with respect to the output labels;
- *numerosity*: whether the number of samples is reasonable with respect to the number of variables;
- *unevenness*: how frequent outliers are;
- *incompleteness*: how frequent missing values are within the dataset;
- *quality*: a manually set feature evaluating the quality of the dataset, with respect to noise or similar characteristics;
- *(legal) compliance*: whether the creator(s) of the dataset respected all laws and legal duties.

FanFAIR aggregates these features into an overall fairness score by using a Fuzzy Inference System (FIS) based on a 0-order Sugeno reasoner. The system relies on fuzzy logic to handle the inherent fuzziness in fairness assessments, which is generally not black-or-white and cannot rely on crisp arbitrary thresholds. FanFAIR is semi-automatic because most of the process is automated, although two metrics (namely, quality and legal compliance) inherently require a human intervention.

4.2.2 New functionalities in FanFAIR

In this work, we updated FanFAIR with three new functionalities: (i) the analysis for sensitive variables; (ii) the identification of outliers with an improved isolation forest able to deal with missing values in data; (iii) extended support for Pandas dataframes and non real-valued variables in the dataset. These functionalities are described in the following subsections.

4.2.2.1 Sensitive variables analysis

FanFAIR exposes a new method to specify which variables of the dataset should be considered sensitive. In this work, we define as "sensitive" all the variables that should not be directly responsible for a given prediction of the system.

The user can now specify a list of sensitive variables using a novel `set_sensitive_variables()` method of the `FanFAIR` object (we will denote by $\mathcal{S}$ the list of sensitive variables). So doing, FanFAIR will automatically calculate the Pearson's correlation coefficient between each input variable $x \in \mathcal{S}$, and the output variable y [1], which is computed as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4.1}$$

where $\bar{x}$ and $\bar{y}$ denote the sample mean of the sensitive variable and the output variable, respectively. Pearson's correlation is symmetric, so that $r_{xy} = r_{yx}$. The possible values of r_{xy} range between -1 and $+1$, representing negative and positive correlation, respectively. Since we are only interested in detecting correlation, we will discard the sign by considering the absolute value, i.e. $R_{xy} = |r_{xy}|$. Finally, we will assess the fairness with respect to sensitive input variables of the dataset as:

$$\rho = \max_{x \in \mathcal{S}}(R_{xy}). \tag{4.2}$$

The rationale of Equation 4.2 is that a high correlation of the output to even a single variable marked as sensitive (e.g., gender, ethnicity, age, political orientation) is enough to classify the whole dataset as unfair. In such a case, the removal of the variable from the dataset before the training of the model is strongly advised.

4.2.2.2 Improved Isolation Forest

We integrated an improved implementation of Isolation Forest [79, 80], which enables FanFAIR to perform multivariate outlier detection leveraging any combination of numeric, boolean, or categorical variables, and also providing support for data with missing

[1] Please note that FanFAIR is currently limited to datasets related to classification problems. Hence, it can be used for regression problems only if the output value is properly converted, e.g., by means of binning.

values. The underlying implementation is provided by the Python module "isotree"[2], which was added to FanFAIR's dependencies. This new method may be used by setting `outliers_detection_method="isotree"` during the creation of a `FanFAIR` object.

Internally, a score between 0 and 1 is computed for each row in the data, with higher scores indicating that the corresponding sample presents a combination of values that is more unusual than those seen in the rest of the data. We then determine the outlier status O_i of each row by performing the following computation:

$$O_i = \begin{cases} \texttt{TRUE} & \text{if } o_i > \min(0.7, \mu_o + 3\sigma_o) \\ \texttt{FALSE} & \text{otherwise} \end{cases} \tag{4.3}$$

where o_i is the score associated to the i-th row, and μ_o and σ_o are the mean and standard deviation of the score across the dataset, respectively. We opted to use the dynamic thresholding detailed in equation (4.3), rather than applying a fixed threshold, to adapt the sensitivity of the method to the sparsity of the available data. Furthermore, we employ the default parameters for the detector (500 trees; 1 randomly selected variable per split; maximum tree depth = height of balanced binary tree with a number of leaves equal to the samples), which we reckon should perform well in most cases.

4.2.2.3 Extended support for datatypes and dataframes

It is now possible to feed a Pandas[81] dataframe directly to FanFAIR through the `dataframe` parameter when creating a `FanFAIR` object. This enables an easier integration of FanFAIR into pipelines that adopt this very popular Python module, and, more crucially, grants the user control over the determination of variables' datatypes, providing an alternative to automatic type inference carried on by the csv parser.

Concurrently, we provided FanFAIR with the ability to automatically select the appropriate variables during each step of the computation of the fairness score, accordingly to the chosen parameters (e.g., during the determination of outliers, binary and categorical features are only preserved when employing "isotree", while dates are excluded in all cases), informing the user of the actions taken. These additions improve FanFAIR's usability, and extend its applicability to a wider variety of datasets.

4.2.3 Case study: COVID dataset

To investigate how data processing applied during the development of an AI system can affect the fairness of a dataset, we relied on our previous study (presented in Chapter 3), as it provided us with several incrementally more processed versions of the same clinical dataset, corresponding to subsequent stages of the data processing pipeline.

[2]The documentation, complete with the list of literature referenced by the developer, is available online on the original author's GitHub repository.

4.2.3.1 Relevant stages of the data processing pipeline

More specifically, we used FanFAIR to evaluate four versions of our dataset:

1. **raw** - (947 rows, 129 columns) imported from csv, it comprises all the available variables; minimal processing was applied, just to allow FanFAIR to evaluate the data, namely: datatype enforcement, removal of patients without outcome, and automated removal of invalid entries of numeric variables;

2. **clean** - (825 rows, 79 columns) uninformative and highly sparse variables were removed, as well as rows pertaining patients not belonging to the population of interest (i.e. those that were not treated with GCs), or presenting clearly anomalous values. Only two out of the nine serially measured variables available were preserved, namely, PaO_2/FiO_2 ratio and CRP level;

3. **reseried** - (825 rows, 57 columns) dates are removed, raw serial measurements are condensed into three fields per series, namely, first and last sample, and a binary "improving" field computed from the raw readings according to heuristics designed by medical experts;

4. **selected** - (825 rows, 10 columns) it only includes the 9 predictors determined during the variable selection phase of the processing pipeline adopted in the original study, and the outcome.

4.2.3.2 Quality, compliance and sensitive variables

In order to evaluate our dataset(s) with FanFAIR, we need to assess the appropriate values for the `quality` and `compliance` parameters, as well as identify sensible variables in our data (as per the definition given in section 4.2.2.1).

Concerning the quality, a fair assessment should be based on the reliability of the data collection and on the professional opinion of domain experts. In our case, the data collection consisted in the manual annotation of the values by clinical doctors over multiple sessions for each patient, and the final manual assembly of the collected data into a single Excel dataset. Human errors might occur at different steps of this articulated process, resulting in noisy data values. As a matter of fact, we discovered evidence of such errors during an extensive quality checking of the date variables in our dataset. The domain experts deliberated that a penalty of 0.1 would be sufficient to account for this noise. Additional source of noise, due to unforeseeable factors, should be accounted too. For these reasons, we decided to use an overall penalty value of 0.2. We consider the final value of 0.8 a very conservative estimate of the quality of the least processed version of our dataset (i.e. raw). In addition, it would be reasonable to assume that subsequent processing steps improved the quality of the data; nonetheless, we deemed it safer to

adopt the same quality value across all the dataset, to prevent any artificial inflation of the final fairness score.

Compliance assessment was rather straightforward, since some of the authors of the present work were directly involved in the original studies that produced the dataset [82, 17]: the clinical and medical data were collected in compliance with legal and ethical standards. Specifically, the data were pseudonymized, handled in accordance with transparency obligations and the rights of the individuals, and the appropriate legal compliance was performed. Formal authorization from the Central Ethical Committee and informed patient consent were obtained prior to data collection. All relevant clinical and diagnostic regulations were followed. Lastly, principles of non-discrimination, fairness, and other standard ethical guidelines were adhered to in the collection, storage, and use of the patient data. On the basis of these considerations, we determined that our dataset meets all five compliance criteria considered by FanFAIR, namely: *data protection*, *copyright*, *medical*, *non discrimination* and *ethics*.

Finally, we identified age and sex (referring to the assigned sex at birth, not the gender identity of the patient) as the only sensible variables in our dataset; the latter in particular was excluded during the variable selection phase of the original study, therefore it does not appear in the final instance of the dataset. It is worth noting that we do expect to detect some degree of correlation between age and mortality in our data, given that the results of the previous study indeed point to age as the second most influential variable, in terms of mean absolute SHAP value associated to the final model.

4.2.3.3 Training and evaluating ML models

To show how model performance and fairness evolve in parallel, we train a Random Forest (RF) classifier on each dataset, and estimate its generalization performances in terms of classic ML metrics. We adopt a very basic pipeline to train and evaluate the models, designed to ensure that both technical and problem-specific prerequisites are met in all four instances of the classification task (briefly summarized in Section 4.2.3.1). Specifically, the pipeline articulates as follows:

1. task-specific selection - patients that did not undergo GCs treatment are excluded (122, only in raw); variables that would make the prediction task trivial, according to our medical experts, are removed (2, in all datasets except selected);

2. technical preprocessing - date columns are dropped; categorical values are replaced with the respective numeric codes; missing values are filled with the median (or modal, in case of categorical) values of the respective variables;

3. model training and evaluation - a vanilla `RandomForestClassifier` (scikit-learn v1.5.1[83]) is trained and evaluated adopting a 5-fold cross-validation scheme with stratification, to preserve outcome proportions.

Hyper-parameter tuning was skipped, as the achievement of optimal performances was not within the scope of the present work. It's also worth noting that, since the variables present in the last dataset (i.e. "selected") were chosen accordingly to a different pipeline, we do not expect that models trained on it according to the present pipeline will necessarily achieve the best overall performance scores.

4.3 Results

We applied FanFAIR to the four versions of the COVID datasets described in Section 4.2.3. We set "age" and "sex" as sensitive variables.

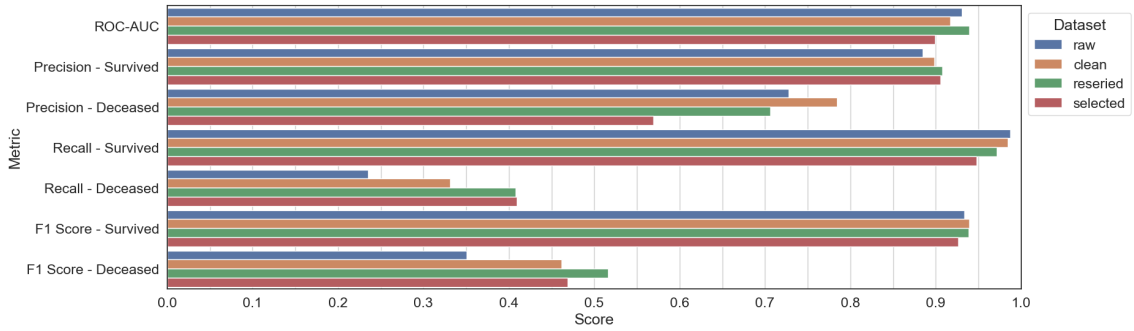

Figure 4.1: Overview of performance metrics of the Random Forest models fitted on each dataset.

The fairness scores calculated by FanFAIR are the following:

- raw : 75.9%
- clean : 82.3%
- reseried : 83.8%
- selected : 84.3%

These values show a clear improving trend in accordance with the progression of the data processing pipeline.

Concurrently, performance metrics (see Figure 4.1) indicate that models trained on more processed versions of the dataset are better at identifying patients who did not survive, which arguably constitutes the most crucial aspect of the task. Specifically, recall on deceased patients using the "selected" dataset almost doubles with respect to the "raw" dataset (from 23.6% to 40.9%), and likewise the F1 score on deceased patients improved by more than 10 percentile points (from 35.1% in "raw" to 46.9% in "selected", and up to 51.7% in "reseried"); the other metrics are only slightly affected, with the sole exception of precision for deceased, which dropped from 72.8% in "raw" to 56.9% in "selected", while still resulting improved with respect to the baseline in "clean" (78.4%).

Given the nature of the task and the strong unbalance in the data, we believe that these results support the hypothesis that adequate pre-processing of the data is not only

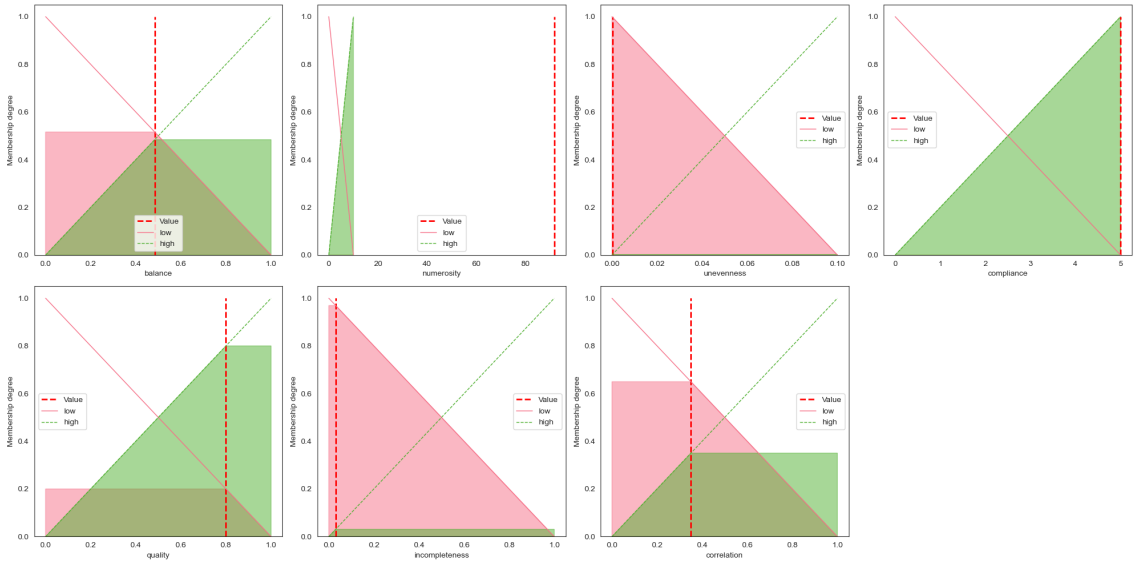

Figure 4.2: Membership values calculated for the fully processed version (i.e. "selected") of the COVID dataset.

necessary to improve the performance of ML solutions, but also constitutes a valid means to reduce the unfairness embedded in the data. In this specific case, the most substantial contribution to the fairness score (+6.4%) was achieved in the first step of the data processing pipeline, which consisted in the exclusion of variables presenting a high level of sparsity (i.e. with 50% or more undefined values), and the restriction of the cohort to patients within the population of interest.

Concerning the analysis of sensitive variables, FanFAIR reports a noteworthy level of correlation between age and mortality (32% in raw, increased to 35% in all other datasets), although domain experts are keen to consider this as a manifestation of the well known "age pattern of mortality", rather than the evidence of the unfair administration of treatments at the expense of elderly patients, especially considering that half of the cohort belongs to this category (median age is 65). The 3% increase in the correlation can be explained by the fact that patients that were dropped from the dataset in the first step lowered the mean age among deceased in the first dataset (Figure 4.3). Furthermore, FanFAIR reports negligible correlation between sex and outcome (2% in "raw", 0% in "clean" and "reseried").

A further inspection of the membership functions reveals that an important issue of the dataset is the balance of the labels, which was assessed around 50% (see Figure 4.2, top-left panel). All other variables do not seem to have a negative impact to the fairness, with the exception of quality which could be slightly improved, e.g., by automating (parts of) the data collection procedure. The numerosity, according to FanFAIR, was excellent to develop a fairer predictive model.

For this dataset, the overall improvement of fairness due to pre-processing was limited ($< 10\%$). Although this may not necessarily be the case for every dataset, this result is in

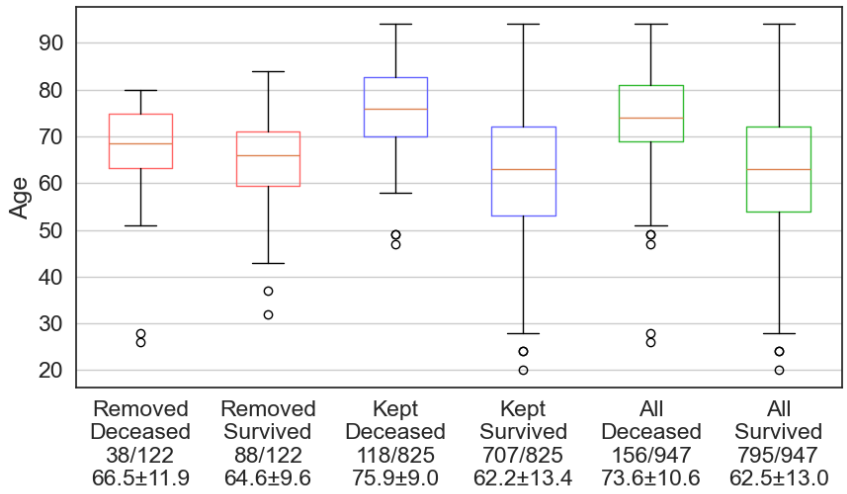

Figure 4.3: Box-plots of age, stratified by outcome, in the group of patients excluded during the first data processing step (left) against the rest (center), and in the entire raw dataset (right); labels indicate numerosity, and age mean±standard deviation for each group.

agreement with the idea, discussed in this work, that the fairness of data (and hence, of the AI systems trained on them) is dependent also on additional factors that must be taken into account since the beginning of the study. These factors include (but are not limited to) a proper design of the data collection phase, and a careful review and fulfillment of all legal requirements that apply to research, at both general and case-specific levels.

4.4 Conclusion

In this work we presented an improvement to FanFAIR, a Python library that leverages fuzzy reasoning for the semi-automatic assessment of datasets fairness. Specifically, we extended FanFAIR to support the specification of one or more sensitive variables. A correlation check is automatically performed on all sensitive variables and that information is later fed to a fuzzy rule, which determines whether the sensitive variables have an excessive influence on the output labels. This preliminary analysis can help to identify potential bias in the data that can be traced back to even a single variable. We also implemented a few missing but useful functionalities, e.g., the import of Pandas' dataframes.

FanFAIR was applied to the analysis of a COVID dataset. Our results showed that both data fairness and model performance can improve in parallel when appropriate data processing is set in place during the development of AI systems. In this regard, FanFAIR can help researchers in deciding whether a given dataset should be pre-processed, or discarded altogether, due to its limitations.

From a computational standpoint, we report that FanFAIR was able to process our tabular dataset in under a minute on a consumer laptop equipped with a CPU Intel i3 of 12^{th} generation, which is arguably efficient for any similar use case. Early testing per-

formed during our study showed that the newly introduced multi-variate outlier detection methods[3] are more computationally demanding than the available uni-variate options, which could be expected. We plan to further inquire on the complexity and scalability of FanFAIR by systematic benchmarking and testing, as this would provide the insight necessary to extend our tool's applicability and reliability.

It is important to highlight that FanFAIR is not a debiasing algorithm, but rather a means to help healthcare workers and data scientists to pre-evaluate the dataset they intend to use to train a ML model, as we demonstrated by applying it to a study case on COVID-19 data. In addition, FanFAIR does not claim to address all potential biases related to AI fairness, as many discriminatory factors are systemic and embedded in society, such as the unavailability of healthcare for certain marginalized groups and their consequent absence in the dataset. Our tool aims to provide a way to assess the dataset so that the risk of unfairness is decreased.

It is often the case that a combination of variables might contribute to a discriminatory prediction. We will extend this feature in future versions of FanFAIR with the possibility to perform multi-variate influence analysis of sensitive variables and also to perform *post-hoc* analysis by leveraging user-provided ML models trained on the data.

FanFAIR is designed to be as intuitive as possible, with a minimal interface, and the capability to autonomously determine the most appropriate values for the parameters that regulate its functioning. The rationale is that our tool is designed to be used by anyone, including professionals who are not experts in the fields of machine learning and data science. Nevertheless, we understand that many practitioners might want to tweak and configure some of its internal settings (e.g., the hyper-parameters of the outlier detection algorithms). As future developments, we will provide FanFAIR with additional (optional) arguments to properly choose such settings.

[3]Benchmarks for the isotree and PyOD Python modules are available at the respective homepages.

5

Spectral Clustering-Powered Survival Analysis for heterogeneous clinical datasets: a case study on COVID-19

The present chapter is based on the submitted manuscript "Spectral Clustering-Powered Survival Analysis for heterogeneous clinical datasets: a case study on COVID-19", of which I am first and corresponding author [9].

5.1 Introduction

Tabular clinical data are most often heterogeneous [84, 85], due to their very nature and to the variety of methods and supports employed for their collection, especially in applied research. Clinical data are notoriously hard to collect [4, 5, 16, 86] and often suffers of a considerable degree of imprecision [87]. These criticalities emerge in most concrete scenarios, such as treatment of rare diseases, clinical response to epidemic outbreaks, or the identification of bio-medically relevant subpopulations of patients affected by major diseases. These problems negatively impact on general health data analytics, as they are the main culprits for the limited size and sub-optimal quality of health datasets [2, 88], limiting the applicability of more data-intensive Machine Learning (ML) techniques.
Furthermore, the existence of several unavoidable constraints on retrospective studies imply that datasets that are small (from the viewpoint of ML) are still going to be widely employed in field health research. In fact, similar problems occur in many other research fields, urging ML researchers to provide targeted methods: indeed, using more sophisti-

cated methods in presence of small heterogeneous datasets may lead to wrong and likely harmful conclusions [89].
The motivation to write this work stemmed from an investigation we conducted on a specific medical dataset which presented the aforementioned criticalities; the analysis of our dataset serves as pilot case study to present a novel and widely applicable technique for the analysis of relatively small survival dataset (i.e. within the thousand of samples). More specifically, our dataset includes 946 hospitalized COVID-19 patients for which we wanted to analyze the survival dynamics. We had previously conducted another study to investigate the applicability of classification ML algorithms on this dataset [7].
In the present study, we have designed and applied an original pipeline which leverages unsupervised clustering[26] and survival analysis[90], with the aim of studying the survival behavior of COVID-19 patients, and how they relate to the available features.

The main goals of this work are the following:

1. providing a replicable pipeline to identify and characterize subpopulations with differing survival behaviors in challenging datasets;

2. analyze our specific dataset with it, concretely demonstrating our method.

We will begin by reviewing the ML and statistical techniques featured in our method (Section 5.2), and then present our pipeline (Section 5.3); Subsequently we present the dataset from our case study, discussing in detail how we applied our method to analyze it, and reviewing the results of the analysis (Section 5.4). We then conclude with a short summary and some remarks on the method's applicability and limitations (Section 5.5).

5.2 Background

This section provides a brief overview of the techniques employed in our method.

5.2.1 Spectral Clustering

Spectral clustering [91, 92] is a powerful and versatile non-parametric clustering method, which belongs to the broader class of unsupervised learning (UL) techniques [26]. It leverages spectral decomposition to partition the dataset, and consists of the following steps:

1. a *similarity measure* is chosen and computed between each pair of points. The resulting $N \times N$ matrix M can be interpreted as the adjacency matrix of a complete weighted graph;

2. the corresponding (normalized) graph laplacian [93] and its eigenvalues are computed;

3. the number of clusters k is determined by identifying the tallest gap in the spectrum of the laplacian;

4. a low-dimensional embedding of the datapoints, corresponding to eigenvectors of the k smallest eigenvalues, is obtained;

5. cluster labels are computed by applying an auxiliary clustering algorithm (e.g., K-Means [26]) to the embeddings.

Crucial choices are the similarity measure, the number of clusters, and the auxiliary clustering algorithm.
The main advantage of this technique is that it is capable of identifying hidden non-linear structures in the data, since it makes no assumption on the cluster shapes. On the other hand, the complexity of the computation scales polynomially with the number of samples, possibly resulting unpractical to process large datasets (e.g., $N > 10'000$), which, however, are not the focus of our study.

5.2.2 Survival Analysis

Survival Analysis (SA)[94, 95, 96, 97, 98] is a statistical framework, based on Continuous-Time Markov Chains [99, 100], for modeling time-to-event and making inferences on them. SA finds its most important and impactful application in medicine, to model time from an initial observation to a specific clinical event of interest. Examples of the latter are patient death, recurrence of a disease (e.g., a cancer recurrence after a chemotherapy), and patient recovery. SA has a key role in several other fields as well, such as engineering, to model time to machine failure [101], business intelligence, to model time until customers churn[102], and several others.
In the context of SA, time is expressed relatively to the individual subject's *time of origin* ($t = 0$), and limited up to a maximum *follow-up time* T. The main subject of inquiry of SA is the *survival probability* $S(t[, x])$, that is, the probability that a subject with features $x = (x_1, ..., x_d)$ will not experience a predetermined *event of interest* (EOI) before time t. Equally important is the *hazard* function $h(t[, x])$, corresponding to the rate at which an individual risks to experience the EOI at time t. The two functions are related by means of the following equation:

$$h(t) = -\frac{d}{dt}[log(S)].$$

A key difficulty in this type of analysis is dealing with *right censoring*, that is, the fact that some patients may have exited the study before experiencing the EOI, which is particularly common in the context of medical studies, e.g. hospitalized patients are dismissed as soon as they successfully recover, before the end of the study, and miss follow-up communication (patients "lost to follow-up"); or the study may have ended while the patient was still hospitalized. For such a patient x, information may only be available up to a given *censoring time* t_x
The *Kaplan-Meier method* (KM)[103] provides a nonparametric estimate of S which relies on observed survival times, both censored and uncensored, according to the following

expression:

$$S(t) = S(t_{j-1})\left(1 - \frac{D_j}{N_j}\right) \quad t \in [t_j, t_{j+1})$$

where $0 < t_1 < t_2 < ... < t_N \leq T$ are the ordered event times of subjects, N_j is the number of subjects under observation just before t_j (that is, those that didn't experience EOI, and weren't lost to follow-up at $t < t_j$), and D_j is the number of subjects who experienced the EOI at t_j. $S(0)$ is set to 1, to reflect the fact that neither subject has experienced EOI before the time of origin. The model relies on the basic assumptions that event times are mutually independent, and that censoring is uninformative. KM does not leverage other features in the data, nonetheless it is possible to compute the KM estimate for different groups of subjects, and compare them [103, 104]. This is commonly accomplished by means of a nonparametric *log-rank* test, which compares the estimated hazard rates of two survival curves [104].

Another fundamental method in survival analysis is *Cox proportional hazard regression* (Cox PH) [98], which – fully taking into account the partial information linked to *right censored* patients – directly models the hazard for each patient as a function of the features in the data, according to the following expression:

$$h(t, x) = h_0(t) \exp(b \cdot x).$$

This model assumes that all subjects in the cohort share a common base hazard $h_0(t)$, which scales proportionally to a term that depends explicitly on a linear combination of the values of the features $(x_1, \ldots, x_d) = x$ with coefficients $(b_1, \ldots, b_d) = b$. The terms $\exp(b_d) > 0$ are known as *hazard ratios*, and offer a clear interpretation of the contribution of each feature: values above 1 suggest the existence of a positive correlation between the corresponding feature's value and the probability of experiencing EOI, while values below 1 indicate a negative correlation.

5.2.3 Landmark Analysis

Falsely assuming feature values are known at baseline introduces the so called *immortal-time bias* [105, 106], which causes over-estimation of beneficial effects, under-estimation of harmful effects, and pretend beneficial effect of factors that are actually unrelated to survival. Moreover, the farther from baseline the information actually becomes available, the larger the bias.

A possible approach to mitigate this important problem is the use of *landmark analysis*[107, 108, 109], which consists in selecting multiple time points, called *landmarks*, and repeating the analysis basing on the information available at those times; in practice, this requires excluding subjects that have experienced EOI or have been censored by the landmark, and updating the values of time-dependent features for the other subjects.

The choice of landmarks should be done basing on when information becomes available (e.g.,in proximity of sampling times adopted during data collection), and, possibly,

leveraging field knowledge to determine when this information is less likely to change, as high variability may still introduce bias.

5.3 Proposed Method

In a nutshell, the key idea behind our method is to combine clustering, which is very effective at extracting complex patterns in the data, with the well established survival analysis framework: the cluster labels act as an hidden feature, condensing the information carried by the original co-variates into a single categorical feature [110, 111, 112]; we propose to compare the KM curves associated to the clusters and to fit a Cox PH model onto the cluster labels at each landmark, in order to determine if clustering managed to capture any pattern that is relevant for survival. In such case, these survival patterns may be characterized by comparing the features' distributions between the clusters using classical statistical tests.
By combining the high versatility of unsupervised learning with more traditional techniques, we aim to offer an interpretable and straightforward alternative to perform multivariate survival analysis with time-varying features, bypassing the limits that affect other available alternatives, such as the need of more conspicuous datasets of deep survival models[113], or the increased difficulty in interpreting results [97].
The procedure to apply our method may be summarized in the following steps:

1. Data Preparation
2. Tuning
3. Fit and Analyze results

In the next sections we will discuss each of these steps in detail.

5.3.1 Data preparation

The aim of this step is assessing and extracting time-dependent information from the data.
First of all, since we intend to perform survival analysis, it is necessary to define and identify fields associated with origin time, event of interest, and censoring, as well as defining the maximum duration of follow-up (i.e. the time after which censoring is implied), so that EOI and censoring times can be obtained for each subject.
Second, time-dependent features must be distinguished from those that remain constant (throughout follow-up); it is necessary to provide a method that leverages the available time data to evaluate time-dependent features at the chosen landmarks. Some examples of how to accomplish this in practice are provided in Table 5.1.

Time-dependent features for which timing data is not available should be excluded (e.g., features signaling that a patient began a treatment at some unspecified time during

Feature(s) in the dataset	Time-dependent feature	Final data-type
Date of event	Event occurrence flag/count	boolean/numeric
Start/End date of condition	Condition duration	numeric
Serial measurements	Earliest/Latest measurement	(same of series)

Table 5.1: Examples of extraction of time-dependent features.

followup). Likewise, samples (i.e. patients in our case) for which time information is not available may only be included in the analysis at baseline (i.e. $t = 0$), and should be excluded at later landmarks. Finally, in this phase we must choose landmark times, basing on field-specific knowledge.
Summarizing, the outputs of this step are:

1. EOI and Censoring times
2. List of landmarks
3. Series of datasets, one for each landmark.

5.3.2 Tuning

This step consists in determining the details necessary to perform clustering.
Our method aims at identifying subpopulations which differ in terms of survival and features distributions, describing how they evolve through follow-up. To ensure comparability, we propose to apply a common clustering pipeline at each landmark. The tuning phase consists in determining the "hyper-parameters" for spectral clustering, which will be fixed for all landmarks. These parameters are:

1. The affinity measure (and any processing necessary to evaluate it)
2. The number of clusters
3. A cluster labeling criteria

We recommend adopting an unbiased affinity measure, that is, one which leverages all the available features and weights their contribution equally (see Section 5.4.2 for a concrete example). Doing otherwise might steer the analysis towards conclusions that are not purely data-driven, although it may provide an opportunity to counterbalance any bias that is demonstrably present in the data, given the existence of solid field knowledge and/or dataset specific considerations in support of more nuanced choices. Moreover, preliminary data cleaning, as well as the feature selection strategy detailed in Section 5.3.3, are meant to minimize the bias due to confounding features.

The number of clusters K should be determined basing on the state of data at the earliest landmark, to avoid any bias deriving from using future information to take decisions in the past. Besides this detail, this operation can be carried out according to the standard procedure, which we briefly outlined in Section 5.2.1.

Finally, it is necessary to define a criteria to name the clusters, since spectral clustering assigns random labels to each of the identified clusters. Unlike the number of clusters, this labeling can leverage future information (e.g., the outcome), as it only affects how the inherently retrospective results will be presented.

5.3.3 Fit and analyze results

This step consists in actually computing the clusters, and applying statistical tests to qualify and measure if and how clusters differ.
To begin, spectral clustering is performed multiple times at each landmark, adopting the parameters defined previously, to carry on the following backwards feature selection procedure: for each feature, clustering is performed with and without it, and the Normalized Mutual Information index (NMI) between the resulting clusterings is measured. The feature corresponding to the least decrease in NMI is then removed, and the process is repeated until removing any feature would cause an excessive drop in NMI. We recommend pre-selecting a sufficiently high stopping threshold (e.g.,0.98), to ensure the least bias, although it would also be possible to observe the variation of NMI as all the features are gradually removed, and then choose the threshold basing on the tallest gap between subsequent estimates. The final spectral clustering is performed on the reduced feature pool, and the chosen labeling criteria is applied to obtain the final clusters.
Clusters at each landmark are then compared, in terms of Kaplan-Meier survival curves, by performing log-rank tests between them, in terms of hazard ratios, obtained by fitting a Cox PH model onto the cluster labels, and in terms of distributions of all features (i.e. regardless of their exclusion during feature selection), by performing the appropriate statistical tests (e.g., Chi-squared for categorical/binary features, and Kolmogorov-Smirnoff for numeric features).
Furthermore, the evaluation of these indicators at each landmark offers a comprehensive description of how the sub-populations characterize in time.

5.4 Case Study

In this section we present a concrete example of how our method can be applied to analyze a clinical dataset. The analysis of this case study constituted the main motivation behind the development of our method.

5.4.1 Dataset

Starting from the raw dataset, presented in section 2.2, we included all the patients for which outcome was known, and time data was present, obtaining a pool of 944 patients. Furthermore, we analyzed 79 potentially relevant features, including medical history, therapy, complications, relevant dates, and serial measurements of *arterial partial pressure of oxygen to fraction of inspired oxygen ratio* (PaO_2/FiO_2, mmHg), and of *C-reactive protein*

levels (CRP, mg/L) .

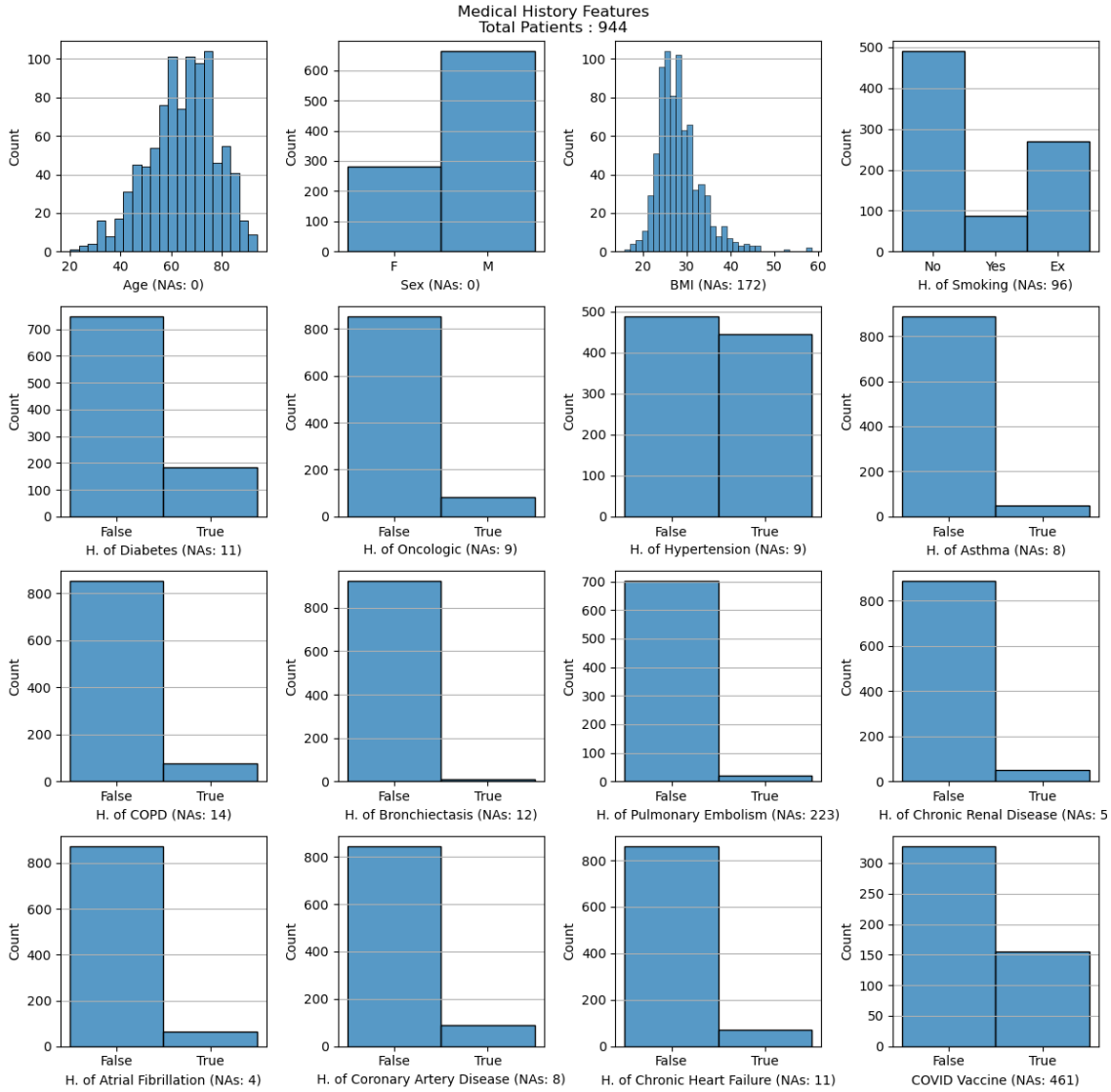

Figure 5.1: Distribution of medical history features in the final dataset, for the population at time 0.

5.4.2 Applying our method

The goal of the present analysis is investigating the cohort's survival times within 60 days from hospitalization. Since part of the data was already cleaned during previous studies, cleaning in this case mostly concerned date fields' validity and consistency with other features; it resulted in the manual fixing of erroneous dates, and the exclusion of 2 patients (deceased at unspecified date).

Decease and censoring times from hospitalization were computed from the respective date fields and capped at 60 days; note that only 14% of the included patients are still hos-

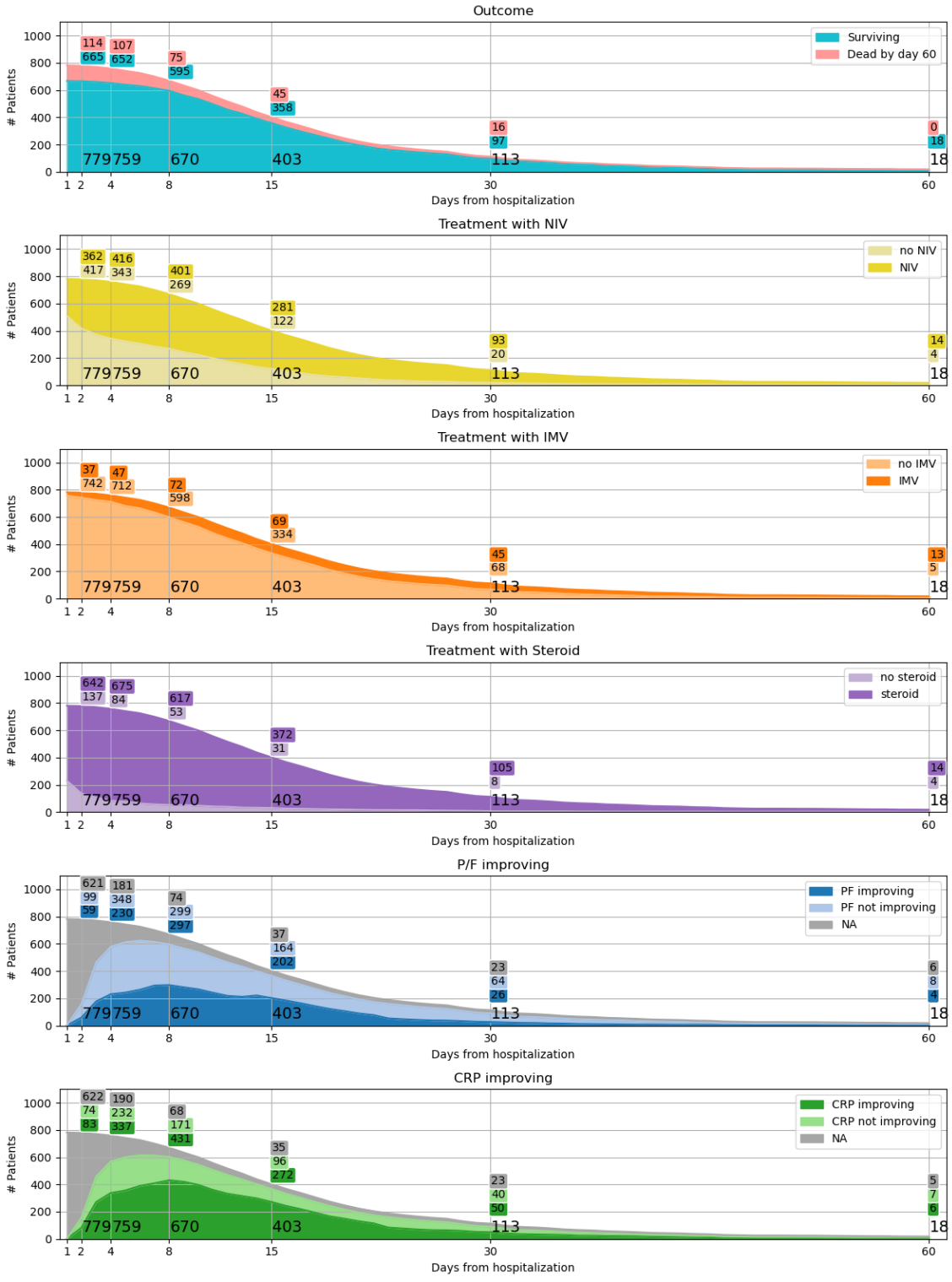

Figure 5.2: Data count (top plot) and evolution of categorical time-varying features in the final dataset. Detailed counts are displayed for landmarks at times $t = 2, 4, 8, 15, 30$ and 60.

pitalized by day 30. In order to properly apply landmark analysis, we had to exclude from the analysis all the features whose values became available at unknown times during

hospitalization, that is, complications and most therapies, whereas we condensed each set of serial measurements into three time-dependent features.

The final datasets comprise 24 features summarized in the following:

- **Medical History** - 15 features - (Figure 5.1) recorded at hospitalization, these features remain constant throughout the follow-up, and include age, biological sex, smoking habits, and health risks tied to the patient's medical history (e.g.,hypertension, asthma, diabetes, etc...);

- **Therapy** - 3 features - (Figure 5.2) detail whether the patient has been treated with non-invasive ventilation (NIV), invasive machine ventilation (IMV), and steroid therapy; these are time dependent, since their value is deduced from the respective therapy start dates,

- **Physiological measurements' summaries** - 6 features - time dependent, these include the earliest and latest available samples of PaO_2/FiO_2 and CRP, along with a categorical "improving" feature (Figure 5.2), indicating whether the corresponding series presents an improving trend or not:

  - PaO_2/FiO_2 is considered improving if the latest measurement is either above 250 mmHg or presents an increase of at least 40% with respect to the earliest sample;
  - CRP is considered improving if the latest measurement is either below 10 mg/L or it is decreased of at least 40% with respect to the earliest sample;

  A dedicated "undefined" label is assigned when less than two samples are available.

Dates necessary to evaluate time-varying features were unavailable for 163 patients, which had to be excluded at landmarks after baseline (i.e. at $t > 0$). Sampling times presented the most obvious candidates for landmarks, since they were determined before data collection by the medical experts at days 0,1,3,7 and 14; in practice, though, sampling occurring after baseline (i.e. for $t > 0$) was often[1] delayed of one day with respect to the intended intervals. This was caused by problems during the data acquisition phase, such as a delayed insertion of the sample data, or the impossibility to take the sample earlier due to the occurrence of medical complications. Such criticalities were particularly common during the COVID-19 pandemic outbreak in Italy, which constituted the environment for the collection of these data, therefore we decided to keep it into account to ensure maximal data availability. We thus set the final landmarks at days 0 (=baseline), 2, 4, 8 and 15 from hospitalization.

We then determined the details of the pipeline to process the data and apply spectral clustering at each landmark, specifically:

[1] i.e. in more than 50% of the cases, as verified by checking their modal values.

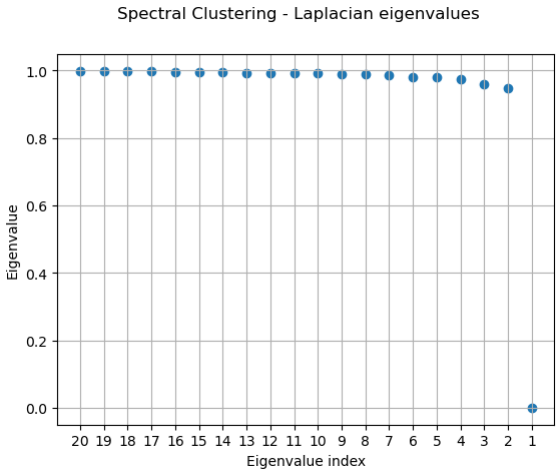

Figure 5.3: Eigenvalues of the graph laplacian for the dataset at $t = 0$, showing that the largest gap occurs between the first and second eigenvalue, and thus suggesting that the appropriate number of clusters is 2.

1. data preprocessing would be applied, to fill missing values in the data (we used the mean value for numeric features, and the modal value for binary and categorical features), and to normalize all feature values to result between 0 and 1. In this case, normalization is preferred over standardization as it ensures that each feature contributes equally to affinity evaluation. Non-binary categorical features in our data (i.e.smoking habits, improving PaO_2/FiO_2 and improving CRP) presented ordered levels, hence we could map them onto evenly spaced values in $[0; 1]$;

2. affinity between patients would be 1 minus the mean absolute difference between the (normalized) values of the patients' features, so that the adjacency matrix would be defined as :
$$M_{i,j} = 1 - \frac{1}{K} \sum_k |x_{i,k} - x_{j,k}| \in [0; 1]$$

3. the number of clusters would be 2, as suggested by the analysis of the eigenvalues of the graph laplacian of our data at time $t = 0$ (Figure 5.3). This implies that we will be fitting a uni-variate Cox PH model and that the covariate can assume two values.

The two clusters computed at each landmark would be labeled *"High Risk"* and *"Low Risk"*, according to the average time-to-decease among included patients.

5.4.3 Results

The results show that the survival curves of the two clusters can be significantly distinguished at each landmark, with high risk patients consistently experiencing an increased hazard (Table 5.2). The greatest separation between the clusters is found at time $t = 8$, in which the hazard ratio between the cluster reaches 4.57: this is consistent with the

Landmark Time	Log-Rank p-value	Hazard Ratio
0	0.00166	1.68
2	2.2e-05	2.28
4	0.000449	2.03
8	**2.98e-05**	**4.57**
15	0.000899	3.94

Table 5.2: Survival curves comparison at each landmark.

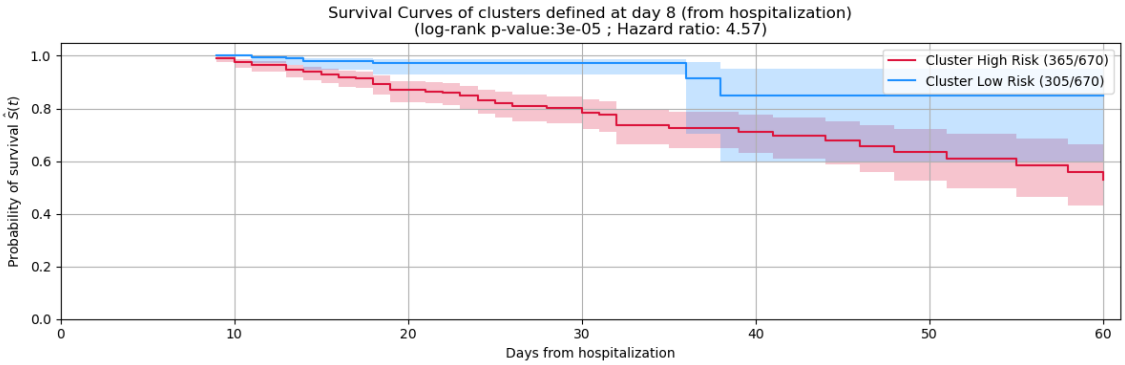

Figure 5.4: Survival curves of the clusters computed basing on the data available at day 8 from hospitalization.

clinicians' experience, who observe that, after a week in hospital, less severe patients have likely been dismissed, while more severe patients are less likely to recover, and, at the same time, more data is available about the patients which are still hospitalized, making it easier to estimate risk.

Feature selection at time $t = 0$ retained only 7 out of the 15 available baseline features, namely: history of Smoking, Diabetes, Hypertension, Chronic Renal Disease, Coronary Artery Disease and Chronic Heart Failure, indicating that these are mainly responsible for the partitioning of the data when therapy information is not yet available; in contrast, fewer features were excluded at other landmarks (six at time $t = 2$, and only one at subsequent landmarks), with all time-varying features contributing to clustering at landmarks $t = 4, 8$, and 15, suggesting their importance in determining the partitioning.

Furthermore, the results detailed in Figure 5.5 provide insights on the role of each feature at each landmark, and across them:

- history of hypertension is always used for clustering, and presents the most separated distribution between clusters at all landmarks (although slightly tapering in time), with higher incidence in the high risk cluster, which is consistent with literature [114];
- age ends among the top five most distinctly distributed features at all landmarks except the last, with a prevalence of older patients in the high risk clusters;
- concordantly with expectations, history of comorbidities result statistically more prevalent in the high risk cluster, with diabetes, coronary artery disease, chronic

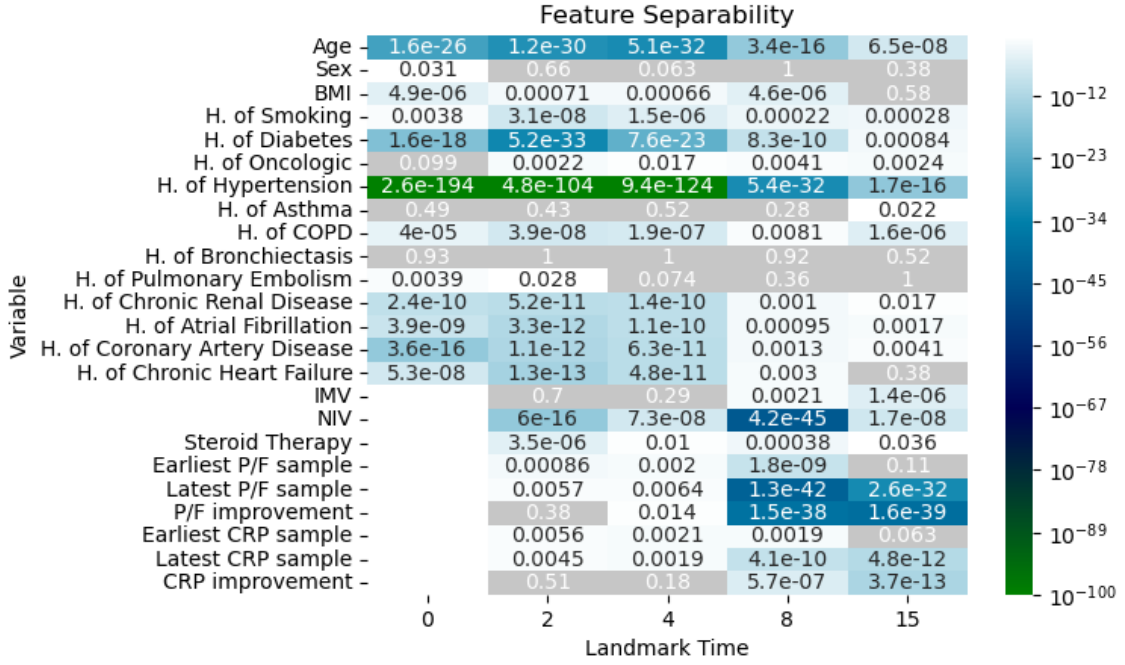

Figure 5.5: Comparison of feature distributions between clusters at each landmark. Cells' label and background color show the p-value of the statistical test used to compare the distributions (Kolmogorov-Smirnoff or Chi-squared); grey background is used when the p-value is above the significance threshold of 0.05. Abbreviations: BMI = body mass index, COPD = chronic obstructive pulmonary disease, IMV = invasive machine ventilation, NIV = non-invasive ventilation, P/F = arterial partial pressure of oxygen to fraction of inspired oxygen ratio, CRP = C-reactive protein.

renal disease, and chronic heart failure presenting the most distinct distributions, although with a degree of significance that slightly decreases with time;

- sex, and history of asthma, bronchiectasis and pulmonary embolism result scarcely separated;

- improvement and latest sample of both PaO_2/FiO_2 and CRP result more distinctly distributed at the last two landmarks ($t = 8$ and 15), with worse values (i.e. low PaO_2/FiO_2 and high CRP) occurring in the high risk group, in agreement with clinicians' expectations;

- Other time-varying features regarding steroid, NIV and IMV therapies result more prevalent in the high risk cluster; in particular, features regarding ventilatory support become increasingly more distinct with time, in concordance with the fact that more severe patients tend to stay in hospital longer and are more likely to receive these treatments.

5.5 Conclusions

In this work we presented an original method which combines unsupervised learning and landmark survival analysis to identify subpopulations in a dataset, describing their survival behavior, and characterizing its relation with the available features, offering a comprehensive view of their evolution in time; furthermore, we demonstrated its application on a medical dataset of 944 COVID-19 patients, successfully partitioning the population into two groups, namely high risk and low risk, which presented statistically distinguishable survival curves and feature distributions.

Our method is designed for the analysis of tabular datasets that include time-varying features, whose limited size (i.e. between hundreds and thousands of rows) could hinder the applicability of more sophisticated methods; in fact, since data scarcity is a frequent condition in applied research for healthcare, as well as other fields, we believe our method is widely applicable, and could provide useful insights to AI practitioners working with real world survival datasets.

We recognize that our work is afflicted by some limitations. First and most importantly, our method relies on the geometric disposition of the data points in the feature space to identify clusters, and thus its effectiveness in splitting the dataset, and the fact that the computed clusters effectively capture differing survival behaviors ultimately depend on the available data. Furthermore, while highly adaptable, our pipeline requires manual intervention of the experts in determining landmark times, and in implementing a method to extract the state of the dataset at the chosen landmark times; we believe this to be inevitable, given the highly case-specific nature of these aspects. Sections 5.3.1 and 5.4.2 provide useful considerations and examples on how to accomplish this.

Finally, while in this work we demonstrated that our method can be used for inference, we believe it could be easily adapted for prediction on larger datasets, by training an additional classification algorithm on the cluster labels computed with our technique.

6 Conclusions

The focus of this work was the development of machine learning (ML) and artificial intelligence (AI) techniques aimed at supporting clinical decision making. To this end, we conducted three complementary studies, each addressing specific methodological and practical challenges.

6.1 Contributions Summary

In the first study we developed a ML pipeline for predicting in-hospital mortality of patients affected by severe COVID-19 pneumonia and treated with glucocorticoids. Multiple supervised algorithms, ranging from logistic regression to ensemble methods such as random forests and extreme gradient boosting, were compared in terms of predictive performance. The best models achieved high discrimination, with AUROC values above 0.9, showing that meaningful predictions can be obtained even from moderately sized datasets when proper preprocessing, feature selection, and validation are employed. Furthermore, explainability was prioritized through the integration of SHAP values, allowing detailed insight into the contribution of each variable to individual predictions. This analysis confirmed the clinical importance of variables such as PaO_2/FiO_2 improvement, C-reactive protein levels, age, and comorbidities, in confirming established medical knowledge and uncovering additional data-driven nuances. The proposed pipeline provides a transparent and versatile tool that could support clinicians in monitoring patients.

In the second study we expanded on the theme of model fairness and data integrity by presenting an improved version of FanFAIR, a hybrid statistical–ML tool designed to assess dataset bias. The enhanced version of the tool features extended data-type support,

compatibility with pandas dataframes, and the possibility to specify sensitive variables, to help identify potential sources of bias. Furthermore, we applied FanFAIR to analyze the COVID-19 dataset, providing evidence that appropriate preprocessing can simultaneously improve predictive accuracy and reduce bias. These findings emphasize the importance of ethical and methodological awareness in ML for healthcare, where biased models may amplify inequities in clinical practice.

In the third study we introduced an original pipeline which combines spectral clustering with landmark survival analysis, enabling the discovery of latent patient subgroups characterized by distinct survival trajectories. This method proved capable of partitioning our COVID-19 dataset into statistically distinct clusters corresponding to low and high risk patient groups, whose survival curves and clinical profiles differed consistently across multiple temporal landmarks. Beyond its specific case study, this approach represents a broadly applicable framework for analyzing survival data in small or heterogeneous datasets where classical models or deep learning approaches may fail due to data limitations.

Together, the three studies illustrate a coherent strategy for leveraging ML in challenging clinical contexts. The combination of supervised learning for outcome prediction, fairness evaluation for bias quantification, and unsupervised learning for subgroup discovery provides a comprehensive toolkit adaptable to a wide range of medical and biomedical datasets.

6.2 Limitations

Nonetheless, several limitations remain. Analyses solely relied on a retrospective dataset, therefore external validation on independent cohorts is required before clinical deployment. Moreover, the interpretability of complex ML systems, though improved through explainable AI techniques, still poses challenges for full clinical integration. Finally, the manual steps foreseen by the presented techniques highlight the need for domain expertise and interdisciplinary collaboration between data scientists and clinicians.

6.3 Future directions

Building on this work, several promising avenues for future research can be identified. First, the generalization and external validation of the proposed models onto larger datasets should be pursued to ensure robustness, reproducibility, and clinical reliability. Integrating these models within EHRs would allow real-time inference, supporting clinicians during decision-making without disrupting established workflows. FanFAIR's functionalities could be extended by adding the possibility to perform multi-variate influence analysis of sensitive variables, as well as post-hoc analysis against user-provided ML models. Furthermore, providing FanFAIR with a stand-alone user interface could greatly improve its usability. The spectral clustering and landmark survival analysis framework could also be

extended to incorporate predictive capabilities, enabling not only subgroup discovery but also dynamic outcome forecasting. Finally, future investigations could explore the controlled use of generative and foundation models, once data availability and transparency safeguards are sufficient, thereby expanding the analytical potential of AI in clinical contexts while maintaining ethical and regulatory compliance.

6.4 Closing remarks

This thesis contributes to the growing body of research bridging ML and clinical practice through methods that are transparent, adaptable, and grounded in real-world constraints. By developing and validating algorithms that respect both statistical rigor and ethical principles, it takes a step toward AI systems that genuinely support the work of clinicians, rather than replace it. The presented approaches form a solid foundation for future interdisciplinary efforts toward explainable, equitable, and effective data-driven healthcare.

Bibliography

[1] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. E. Gabriel, and J. L. Warner, "The evolving use of electronic health records (ehr) for research," in *Seminars in radiation oncology*, vol. 29, no. 4. Elsevier, 2019, pp. 354–361.

[2] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Computer science review*, vol. 40, p. 100370, 2021.

[3] A. F. Simpao, L. M. Ahumada, J. A. Gálvez, and M. A. Rehman, "A review of analytics and clinical informatics in health care," *Journal of medical systems*, vol. 38, no. 4, p. 45, 2014.

[4] M. N. Sarkies, K.-A. Bowles, E. Skinner, D. Mitchell, R. Haas, M. Ho, K. Salter, K. May, D. Markham, L. O'Brien *et al.*, "Data collection methods in health services research," *Applied clinical informatics*, vol. 6, no. 01, pp. 96–109, 2015.

[5] W. R. Hogan and M. M. Wagner, "Accuracy of data in computer-based patient records," *Journal of the American Medical Informatics Association*, vol. 4, no. 5, pp. 342–355, 1997.

[6] L. Pantanowitz, T. Pearce, I. Abukhiran, M. Hanna, S. Wheeler, T. R. Soong, A. P. Tafti, J. Pantanowitz, M. Y. Lu, F. Mahmood *et al.*, "Non-generative artificial intelligence (ai) in medicine: advancements and applications in supervised and unsupervised machine learning," *Modern Pathology*, p. 100680, 2024.

[7] 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*, vol. 28, no. 9, pp. 439–445, 2024.

[8] M. Rispoli, M. S. Nobile, L. Manzoni, A. D'Onofrio, M. Confalonieri, F. Salton, P. Confalonieri, B. Ruaro, and C. Gallese, "Investigating fairness with fanfair: is pre-processing useful only for performances?" in *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*. IEEE, 2025, pp. 1–7.

[9] M. Rispoli, F. Salton, A. Rodriguez-Garcia, A. Rocca, S. Crosera, P. Confalonieri, M. Confalonieri, A. d'Onofrio, and L. Manzoni, "Spectral clustering-powered sur-

vival analysis for heterogeneous clinical datasets: a case study on covid-19," in *submitted manuscript*, 2025.

[10] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New England journal of medicine*, vol. 382, no. 8, pp. 727–733, 2020.

[11] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini, "The covid-19 pandemic," *Critical reviews in clinical laboratory sciences*, vol. 57, no. 6, pp. 365–388, 2020.

[12] W. H. Organization *et al.*, "Who-convened global study of origins of sars-cov-2: China part," 2021.

[13] S. Samson, E. Lord, and V. Makarenkov, "Assessing the emergence time of sars-cov-2 zoonotic spillover," *Plos one*, vol. 19, no. 4, p. e0301195, 2024.

[14] R. Sarker, A. Roknuzzaman, Nazmunnahar, M. Shahriar, M. J. Hossain, and M. R. Islam, "The who has declared the end of pandemic phase of covid-19: Way to come back in the normal life," *Health science reports*, vol. 6, no. 9, p. e1544, 2023.

[15] C. Indolfi and C. Spaccarotella, "The outbreak of covid-19 in italy: fighting the pandemic," pp. 1414–1418, 2020.

[16] K. L. Easton, J. F. McComish, and R. Greenberg, "Avoiding common pitfalls in qualitative data collection and transcription," *Qualitative health research*, vol. 10, no. 5, pp. 703–707, 2000.

[17] 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*, vol. 7, no. 10. Oxford University Press US, 2020, p. ofaa421.

[18] F. Salton, P. Confalonieri, C. Torregiani, B. Ruaro, and M. Confalonieri, "Higher, but not too high, dose is only one determinant of corticosteroid treatment success in severe covid-19," *Annals of the American Thoracic Society*, vol. 20, no. 9, pp. 1371–1371, 2023.

[19] A. Leha, K. Hellenkamp, B. Unsöld, S. Mushemi-Blake, A. M. Shah, G. Hasenfuß, and T. Seidler, "A machine learning approach for the prediction of pulmonary hypertension," *PloS one*, vol. 14, no. 10, p. e0224453, 2019.

[20] C. Meiring, A. Dixit, S. Harris, N. S. MacCallum, D. A. Brealey, P. J. Watkinson, A. Jones, S. Ashworth, R. Beale, S. J. Brett *et al.*, "Optimal intensive care outcome prediction over time using machine learning," *PloS one*, vol. 13, no. 11, p. e0206862, 2018.

[21] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[22] A. Tariq, S. Purkayastha, G. P. Padmanaban, E. Krupinski, H. Trivedi, I. Banerjee, and J. W. Gichoya, "Current clinical applications of artificial intelligence in radiology and their best supporting evidence," *Journal of the American College of Radiology*, vol. 17, no. 11, pp. 1371–1381, 2020.

[23] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[24] S. Marletta, A. Eccher, F. M. Martelli, N. Santonicco, I. Girolami, A. Scarpa, F. Pagni, V. L'Imperio, L. Pantanowitz, S. Gobbo *et al.*, "Artificial intelligence–based algorithms for the diagnosis of prostate cancer: A systematic review," *American Journal of Clinical Pathology*, vol. 161, no. 6, pp. 526–534, 2024.

[25] M. A. Muzammil, S. Javid, A. K. Afridi, R. Siddineni, M. Shahabi, M. Haseeb, F. Fariha, S. Kumar, S. Zaveri, and A. J. Nashwan, "Artificial intelligence-enhanced electrocardiography for accurate diagnosis and management of cardiovascular diseases," *Journal of electrocardiology*, vol. 83, pp. 30–40, 2024.

[26] K. P. Murphy, *Probabilistic machine learning: an introduction*. MIT press, 2022.

[27] Q. Qi, Y. Luo, Z. Xu, S. Ji, and T. Yang, "Stochastic optimization of areas under precision-recall curves with provable convergence," *Advances in neural information processing systems*, vol. 34, pp. 1752–1765, 2021.

[28] J. Wu, Y. Cui, X. Sun, G. Cao, B. Li, D. M. Ikeda, A. W. Kurian, and R. Li, "Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways," *Clinical Cancer Research*, vol. 23, no. 13, pp. 3334–3342, 2017.

[29] M. M. Islam, S. Huang, R. Ajwad, C. Chi, Y. Wang, and P. Hu, "An integrative deep learning framework for classifying molecular subtypes of breast cancer," *Computational and structural biotechnology journal*, vol. 18, pp. 2185–2199, 2020.

[30] A. Roohi, K. Faust, U. Djuric, and P. Diamandis, "Unsupervised machine learning in pathology: the next frontier," *Surgical Pathology Clinics*, vol. 13, no. 2, pp. 349–358, 2020.

[31] T. J. Loftus, B. Shickel, J. A. Balch, P. J. Tighe, K. L. Abbott, B. Fazzone, E. M. Anderson, J. Rozowsky, T. Ozrazgat-Baslanti, Y. Ren *et al.*, "Phenotype clustering in health care: a narrative review for clinicians," *Frontiers in artificial intelligence*, vol. 5, p. 842306, 2022.

[32] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[33] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[34] A. Waqas, M. M. Bui, E. F. Glassy, I. El Naqa, P. Borkowski, A. A. Borkowski, and G. Rasool, "Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models," *Laboratory investigation*, vol. 103, no. 11, p. 100255, 2023.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[36] M. E. Tschuchnig, G. J. Oostingh, and M. Gadermayr, "Generative adversarial networks in digital pathology: a survey on trends and future potential," *Patterns*, vol. 1, no. 6, 2020.

[37] K. S. Chan and N. Zary, "Applications and challenges of implementing artificial intelligence in medical education: integrative review," *JMIR medical education*, vol. 5, no. 1, p. e13930, 2019.

[38] S. Pan, E. Abouei, J. Wynne, C.-W. Chang, T. Wang, R. L. Qiu, Y. Li, J. Peng, J. Roper, P. Patel *et al.*, "Synthetic ct generation from mri using 3d transformer-based denoising diffusion model," *Medical Physics*, vol. 51, no. 4, pp. 2538–2548, 2024.

[39] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, N. Tomasev *et al.*, "Towards conversational diagnostic ai," *arXiv preprint arXiv:2401.05654*, 2024.

[40] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review," *Diagnostic pathology*, vol. 19, no. 1, p. 43, 2024.

[41] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[42] H. H. Rashidi, J. Pantanowitz, M. G. Hanna, A. P. Tafti, P. Sanghani, A. Buchinsky, B. Fennell, M. Deebajah, S. Wheeler, T. Pearce *et al.*, "Introduction to artificial

intelligence and machine learning in pathology and medicine: generative and non-generative artificial intelligence basics," *Modern Pathology*, vol. 38, no. 4, p. 100688, 2025.

[43] A. H. Attaway, R. G. Scheraga, A. Bhimraj, M. Biehl, and U. Hatipoğlu, "Severe covid-19 pneumonia: pathogenesis and clinical management," *bmj*, vol. 372, 2021.

[44] R. C. Group, "Dexamethasone in hospitalized patients with covid-19," *New England journal of medicine*, vol. 384, no. 8, pp. 693–704, 2021.

[45] W. S. Lim, M. M. Van der Eerden, R. Laing, W. G. Boersma, N. Karalus, G. I. Town, S. Lewis, and J. Macfarlane, "Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study," *Thorax*, vol. 58, no. 5, pp. 377–382, 2003.

[46] M. J. Fine, D. E. Singer, B. H. Hanusa, J. R. Lave, and W. N. Kapoor, "Validation of a pneumonia prognostic index using the medisgroups comparative hospital database," *The American journal of medicine*, vol. 94, no. 2, pp. 153–159, 1993.

[47] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, p. 455, 2020.

[48] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, E. Albu, B. Arshi, V. Bellou, M. M. Bonten *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *bmj*, vol. 369, 2020.

[49] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," *Nature reviews Molecular cell biology*, vol. 23, no. 1, pp. 40–55, 2022.

[50] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

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

[52] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification: thinking with examples for effective learning.* Springer, 2016, pp. 207–235.

[53] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big data analytics*, vol. 1, no. 1, p. 9, 2016.

[54] A. Karthikeyan, A. Garg, P. Vinod, and U. D. Priyakumar, "Machine learning based clinical decision support system for early covid-19 mortality prediction," *Frontiers in public health*, vol. 9, p. 626697, 2021.

[55] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[57] N. Sharma, "Xgboost. the extreme gradient boosting for mining applications," 2018.

[58] J. Smit, P. Van Der Zee, S. Stoof, M. Van Genderen, D. Snijders, W. Boersma, P. Confalonieri, F. Salton, M. Confalonieri, M. Shih *et al.*, "Predicting individualized treatment effects of corticosteroids in community-acquired-pneumonia: a data-driven analysis of randomized controlled trials," *medRxiv*, pp. 2023–10, 2023.

[59] Y. Gao, G.-Y. Cai, W. Fang, H.-Y. Li, S.-Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P.-F. Cui *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for covid-19," *Nature communications*, vol. 11, no. 1, p. 5033, 2020.

[60] J. Smit, J. Krijthe, H. Endeman, A. Tintu, Y. de Rijke, D. Gommers, O. Cremer, R. Bosman, S. Rigter, E.-J. Wils *et al.*, "Dynamic prediction of mortality in covid-19 patients in the intensive care unit: A retrospective multi-center cohort study," *Intelligence-based medicine*, vol. 6, p. 100071, 2022.

[61] H. Burdick, C. Lam, S. Mataraso, A. Siefkas, G. Braden, R. P. Dellinger, A. McCoy, J.-L. Vincent, A. Green-Saxena, G. Barnes *et al.*, "Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial," *Computers in biology and medicine*, vol. 124, p. 103949, 2020.

[62] D. Assaf, Y. Gutman, Y. Neuman, G. Segal, S. Amit, S. Gefen-Halevi, N. Shilo, A. Epstein, R. Mor-Cohen, A. Biber *et al.*, "Utilization of machine-learning models to accurately predict the risk for critical covid-19," *Internal and emergency medicine*, vol. 15, no. 8, pp. 1435–1443, 2020.

[63] C. Cilloniz, L. Ward, M. L. Mogensen, J. M. Pericàs, R. Méndez, A. Gabarrús, M. Ferrer, C. Garcia-Vidal, R. Menendez, and A. Torres, "Machine-learning model for mortality prediction in patients with community-acquired pneumonia: development and validation study," *Chest*, vol. 163, no. 1, pp. 77–88, 2023.

[64] L. I. Veldhuis, N. J. Woittiez, P. W. Nanayakkara, and J. Ludikhuize, "Artificial intelligence for the prediction of in-hospital clinical deterioration: a systematic review," *Critical care explorations*, vol. 4, no. 9, p. e0744, 2022.

[65] G. U. Meduri, D. Annane, M. Confalonieri, G. P. Chrousos, B. Rochwerg, A. Busby, B. Ruaro, and B. Meibohm, "Pharmacological principles guiding prolonged gluco-

corticoid treatment in ards," *Intensive care medicine*, vol. 46, no. 12, pp. 2284–2296, 2020.

[66] D. Chaudhuri, A. M. Nei, B. Rochwerg, R. A. Balk, K. Asehnoune, R. Cadena, J. A. Carcillo, R. Correa, K. Drover, A. M. Esper *et al.*, "2024 focused update: guidelines on use of corticosteroids in sepsis, acute respiratory distress syndrome, and community-acquired pneumonia," *Critical care medicine*, vol. 52, no. 5, pp. e219–e233, 2024.

[67] C. Sessa, C. Gallese, F. Schettini, D. Bellavia, F. Asperti, and E. Falletti, "Identifying bias in data collection: A case study on drugs distribution," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–10.

[68] A. Balayn and S. Gürses, "Beyond debiasing: Regulating AI and its inequalities," *EDRi report*, 2021.

[69] S. Milan and E. Treré, "Big data from the south (s): Beyond data universalism," *Television & New Media*, vol. 20, no. 4, pp. 319–335, 2019.

[70] T. Scantamburlo, P. Falcarin, A. Veneri, A. Fabris, C. Gallese, V. Billa, F. Rotolo, and F. Marcuzzi, "Software systems compliance with the AI Act: Lessons learned from an international challenge," in *Proceedings of the 2nd International Workshop on Responsible AI Engineering*, 2024, pp. 44–51.

[71] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 501–512.

[72] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1–20, 2017.

[73] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang, "Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?" in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 325–336.

[74] D. Thakkar, A. Ismail, P. Kumar, A. Hanna, N. Sambasivan, and N. Kumar, "When is machine learning data good?: Valuing in public health datafication," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–16.

[75] C. Gallese, C. Fuchs, S. G. Riva, E. Foglia, F. Schettini, L. Ferrario, E. Falletti, and M. S. Nobile, "Predicting and characterizing legal claims of hospitals with computational intelligence: the legal and ethical implications," in *2022 IEEE Conference on*

Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, 2022, pp. 1–9.

[76] C. Gallese, "Legal aspects of AI in the biomedical field. the role of interpretable models," *Big Data Analysis and Artificial Intelligence for Medical Sciences*, 2024.

[77] E. R. Goffi, L. Colin, and S. Belouali, "Ethical Assessment of AI Cannot Ignore Cultural Pluralism: A Call for Broader Perspective on AI Ethic," *Arribat-International Journal of Human Rights Published by CNDH Morocco*, vol. 1, no. 2, pp. 151–175, 2021.

[78] C. Gallese, T. Scantamburlo, L. Manzoni, and M. S. Nobile, "Investigating semi-automatic assessment of data sets fairness by means of fuzzy logic," in *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2023, pp. 1–10.

[79] W. S. Al Farizi, I. Hidayah, and M. N. Rizal, "Isolation forest based anomaly detection: A systematic literature review," in *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*. IEEE, 2021, pp. 118–122.

[80] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended isolation forest," *IEEE transactions on knowledge and data engineering*, vol. 33, no. 4, pp. 1479–1489, 2019.

[81] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[82] 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*, vol. 61, no. 4, 2023.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[84] J. A. Maldonado, M. Marcos, J. T. Fernández-Breis, E. Parcero, D. Boscá, M. del Carmen Legaz-García, B. Martínez-Salvador, and M. Robles, "A platform for exploration into chaining of web services for clinical data transformation and reasoning," in *AMIA Annual Symposium Proceedings*, vol. 2016, 2017, p. 854.

[85] B. Elsharkawy, H. Ahmed, and R. Salem, "Semantic-based approach for solving the heterogeneity of clinical data," *IJCI. International Journal of Computers and Information*, vol. 5, no. 1, pp. 35–45, 2016.

[86] M. Zlowodzki, A. Jönsson, and M. Bhandari, "Common pitfalls in the conduct of clinical research," *Medical Principles and Practice*, vol. 15, no. 1, pp. 1–8, 2005.

[87] E. Capobianco, "Imprecise data and their impact on translational research in medicine," *Frontiers in Medicine*, vol. 7, p. 82, 2020.

[88] L. Pantanowitz, T. Pearce, I. Abukhiran, M. Hanna, S. Wheeler, T. R. Soong, A. P. Tafti, J. Pantanowitz, M. Y. Lu, F. Mahmood *et al.*, "Nongenerative artificial intelligence in medicine: advancements and applications in supervised and unsupervised machine learning," *Modern Pathology*, vol. 38, no. 3, p. 100680, 2025.

[89] P. Kokol, M. Kokol, and S. Zagoranski, "Machine learning on small size samples: A synthetic knowledge synthesis," *Science Progress*, vol. 105, no. 1, p. 00368504211029777, 2022.

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

[91] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[92] L. Ding, C. Li, D. Jin, and S. Ding, "Survey of spectral clustering based on graph theory," *Pattern Recognition*, p. 110366, 2024.

[93] M. Newman, *Networks*. Oxford University Press, 2018.

[94] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part i: basic concepts and first analyses," *British journal of cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[95] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, "Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods," *British journal of cancer*, vol. 89, no. 3, pp. 431–436, 2003.

[96] ——, "Survival analysis part iii: multivariate data analysis–choosing a model and assessing its adequacy and fit," *British journal of cancer*, vol. 89, no. 4, pp. 605–611, 2003.

[97] T. G. Clark, M. J. Bradburn, S. B. Love, and D. Altman, "Survival analysis part iv: further concepts and methods in survival analysis," *British journal of cancer*, vol. 89, no. 5, pp. 781–786, 2003.

[98] D. R. Cox, *Analysis of survival data*. Chapman and Hall, 2018.

[99] O. O. Aalen, P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, "Martingales in survival analysis," in *The splendors and miseries of martingales: their history from the casino to mathematics*. Springer, 2022, pp. 295–320.

[100] D. Bakry, R. D. Gill, S. A. Molchanov, and R. D. Gill, *Lectures on survival analysis*. Springer, 1994.

[101] S. Softic and B. Hrnjica, "Case studies of survival analysis for predictive maintenance in manufacturing," *International Journal of Industrial Engineering and Management*, vol. 15, no. 4, pp. 320–337, 2024.

[102] O. Çelik and U. O. Osmanoglu, "Comparing to techniques used in customer churn analysis," *Journal of Multidisciplinary Developments*, vol. 4, no. 1, pp. 30–38, 2019.

[103] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding kaplan-meier curves," *Otolaryngology—Head and Neck Surgery*, vol. 143, no. 3, pp. 331–336, 2010.

[104] P. Sedgwick and K. Joekes, "Kaplan-meier survival curves: interpretation and communication of risk," *British Medical Journal*, vol. 347, 2013.

[105] S. Suissa, "Immortal time bias in pharmacoepidemiology," *American journal of epidemiology*, vol. 167, no. 4, pp. 492–499, 2008.

[106] K. Yadav and R. J. Lewis, "Immortal time bias in observational studies," *Journal of American Medical Association*, vol. 325, no. 7, pp. 686–687, 2021.

[107] U. Dafni, "Landmark analysis at the 25-year landmark point," *Circulation: Cardiovascular Quality and Outcomes*, vol. 4, no. 3, pp. 363–371, 2011.

[108] A. Gleiss, R. Oberbauer, and G. Heinze, "An unjustified benefit: immortal time bias in the analysis of time-dependent events," *Transplant international*, vol. 31, no. 2, pp. 125–130, 2018.

[109] C. J. Morgan, "Landmark analysis: a primer," *Journal of Nuclear Cardiology*, vol. 26, pp. 391–393, 2019.

[110] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[111] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen *et al.*, "Molecular portraits of human breast tumours," *nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[112] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the national academy of sciences*, vol. 100, no. 14, pp. 8418–8423, 2003.

[113] L. Xu and C. Guo, "Coxnam: An interpretable deep survival analysis model," *Expert Systems with Applications*, vol. 227, p. 120218, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423007200

[114] M. Peng, J. He, Y. Xue, X. Yang, S. Liu, and Z. Gong, "Role of hypertension on the severity of covid-19: A review," *Journal of cardiovascular pharmacology*, vol. 78, no. 5, pp. e648–e655, 2021.