Developing Personalized Prediction Models for Acute Respiratory Distress Syndrome in the Intensive Care Unit

---

Entwicklung personalisierter Prognosemodelle für akutes Lungenversagen auf der Intensivstation

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Richard Polzin

Berichter: Universitätsprofessor Dr. rer. nat. Andreas Schuppert

Universitätsprofessor Dr. sc. Sebastian Trimpe

Tag der mündlichen Prüfung: 25. September 2024

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Abstract

More and more data is generated every day around the world. The clinical landscape has significantly shifted in recent years, with information increasingly being processed digitally. Especially in the intensive care unit (ICU) of hospitals, sophisticated machines and medical devices produce a plethora of data. Artificial Intelligence (AI) and Machine Learning (ML) can be used to analyze these vast and heterogeneous data spaces in search of underlying patterns and hidden signals, which have so far eluded any attempts of manual analysis and extraction through human experts. Acute Respiratory Distress Syndrome (ARDS) is a life-threatening condition that affects millions of people worldwide, resulting in high morbidity and mortality [1] [2]. During ARDS a widespread inflammation of the lung impairs its ability to oxygenate the body. The syndrome is often under diagnosed and frequently detected delayed, which has a significant impact on the associated mortality [3] [4] [5] [6]. This work presents a prediction model to improve outcomes in ARDS patients in the ICU, based on a new platform for the development of ML models on heterogeneous timeseries data. The model predicts a stark decrease in lung function, as indicated by a loss of oxygenation in the bloodstream. Timeseries data from ICU wards of a German university hospital [7] was used, with features of the past 24 hours summarized through descriptive statistics in a moving window. A gradient-boosted tree ensemble was found to enable the best performance in predicting the probability of a rapid decline in the ratio of the pressure of oxygen in the blood and the fraction of oxygen in the inhaled air. The developed model demonstrates high predictive performance achieving a ROC-AUC of 0.95 outperforming the state-of-the-art models significantly, with a sensitivity and specificity of 0.87 and 0.91 respectively and a three-day prediction horizon. The generalizability of this model was further evaluated, by deploying a model trained on a subset of patients without a COVID-19 infection to predict patients infected with the virus. We show that the model generalizes well in this case and significantly outperforms a model trained on only the infected patients. To summarize, this work contributes a novel framework for the development of ML models on heterogeneous medical timeseries data, which provides a missing piece of software at the intersection of healthcare and research, with a focus on high-performance computing (HPC) and visualization. Based on this framework, a model for the prediction of rapid loss of oxygenation in ICU patients was developed, providing superiority in prediction compared to similar approaches [8] [9] [10], achieving high predictive performance at a large prediction horizon.

Zusammenfassung

Jeden Tag werden auf der ganzen Welt mehr und mehr Daten erzeugt. Die klinische Landschaft hat sich in den letzten Jahren stark verändert, da Informationen zunehmend digital verarbeitet werden. Vor allem auf der Intensivstation (ICU) von Krankenhäusern produzieren hochentwickelte Maschinen und medizinische Geräte eine Fülle von Daten. Künstliche Intelligenz (KI) und maschinelles Lernen (ML) können eingesetzt werden, um diese riesigen und heterogenen Datenräume auf der Suche nach zugrundeliegenden Mustern und verborgenen Signalen zu analysieren, die sich bisher allen Versuchen der manuellen Analyse und Extraktion durch menschliche Experten entzogen haben. Das akute Atemnotsyndrom (Acute Respiratory Distress Syndrome, ARDS) ist eine lebensbedrohliche Erkrankung, von der weltweit Millionen von Menschen betroffen sind und die zu einer hohen Morbidität und Mortalität führt [1] [2]. Beim ARDS beeinträchtigt eine Entzündung der Lunge deren Fähigkeit, den Körper mit Sauerstoff zu versorgen. Das Syndrom wird oft unterdiagnostiziert und häufig verzögert erkannt, was einen erheblichen Einfluss auf die damit verbundene Mortalität hat [3] [4] [5] [6]. In dieser Arbeit wird ein Prognosemodell zur Verbesserung der Behandlungsergebnisse bei ARDS-Patienten auf der Intensivstation vorgestellt, das auf einer neuen Plattform für die Entwicklung von ML-Modellen auf heterogenen Zeitreihendaten basiert. Das Modell sagt eine starke Abnahme der Lungenfunktion, definiert als ein Verlust der Sauerstoffzufuhr im Blutkreislauf, voraus. Für die Entwicklung wurden Zeitreihendaten von Intensivstationen eines deutschen Universitätsklinikums [7] verwendet, wobei die Patientendaten der letzten 24 Stunden durch deskriptive Statistiken in einem gleitenden Fenster zusammengefasst wurden. Es wurde festgestellt, dass ein Gradient-Boosted Tree Ensemble die beste Leistung bei der Vorhersage eines rapiden Abfalls des Verhältnisses zwischen dem Sauerstoffdruck im Blut und dem Sauerstoffanteil in der eingeatmeten Luft ermöglicht. Das entwickelte Modell weist eine hohe Vorhersageleistung auf und übertrifft mit einem ROC-AUC von 0,95 und einer Sensitivität und Spezifität von 0,87 bzw. 0,91, so wie einem Vorhersagehorizont von drei Tagen, den aktuellen Stand der Technik deutlich.. Die Generalisierbarket dieses Modells wurde weiter evaluiert, indem ein Modell, das auf einer Teilmenge von Patienten ohne COVID-19-Infektion trainiert wurde, zur Vorhersage von mit dem Virus infizierten Patienten eingesetzt wurde. Wir zeigen, dass das Modell in diesem Fall gut funktioniert und ein Modell, das nur auf infizierten Patienten trainiert wurde, deutlich übertrifft. Zusammenfassend präsentiert diese Arbeit eine neuartige Plattform für die Entwicklung von ML-Modellen auf heterogenen medizinischen Zeitreihendaten, welche ein fehlendes Stück Software an der Schnittstelle zwischen Gesundheitswesen und Forschung, mit einem Schwerpunkt auf High-Performance-Computing (HPC) und Visualisierung, darstellt. Auf der Grundlage dieses Frameworks wurde ein Modell für die Vorhersage des schnellen Verlusts der Sauerstoffversorgung bei Intensivpatienten entwickelt, welches im Vergleich zu ähnlichen Ansätzen [8] [9] [10] eine überlegene Vorhersageleistung bei einem großen Vorhersagehorizont bietet.

Acknowledgements

First and foremost, I would like to express my gratitude to Prof. Andreas Schuppert for his supervision and guidance. I deeply appreciate the many fruitful discussions and the scientific, curious, and open mindset fostered within our group. I would also like to thank Prof. Sebastian Trimpe for serving as the second examiner of this thesis.

Next, I would like to extend my gratitude to my colleagues at JRC. Most notably, I want to thank Hülya Ulu-Esser for her endless compassion and support—be it organizational, emotional, or otherwise. Thank you for the myriad ways you ensured everything ran smoothly, all while always offering an open ear. I also want to express my thanks to Dr. Konstantin Sharafutdinov, with whom I have worked extensively over the past few years. I thoroughly enjoyed our many discussions, both scientific and otherwise, and I am grateful to have shared so much of this journey with you. I apologize for corrupting your coffee standards—you're welcome. A special thanks to Dr. Nina Kusch for the enjoyable and recreational MATLAB-teaching sessions, as well as the various cat-related highlights we've shared. Thank you to Dr. Jeyashree Krishnan for the many pleasant chats and the ongoing encouragement to travel. I am happy to report that I have finally stepped foot outside of Germany and plan to do so again in the future. During the final stretch of this work, the unwavering support of Jorge Guzman was invaluable to me. Thank you for the plant updates, the late-night talks at the library, and for being a personal motivator. Of course, my gratitude extends to everyone else at JRC as well. Thank you all for creating such a pleasant environment, engaging in interesting discussions, and contributing to the many fun memories.

Within the context of the ASIC project, I would like to thank Dr. Chadi Barakat, Dr. Sebastian

Fritsch, Dr. Johannes Bickenbach, Dr. Gernot Marx, Dr. Joyce Kao, Simon Fonck, Christoph Müller, and Andreas Bleilevens. It has been a true pleasure working with you over the past years, and I will fondly remember the interesting journey we shared.

Finally, I want to extend my gratitude to my friends and family.

Words can not express my gratitude to Chris. Their continued support was a cornerstone of my ability to complete this work, and, on many occasions, they were the golden thread I clung to during challenging times. Thank you for providing a sanctuary and for taking me on countless braincations to relax and recharge. I eagerly look forward to exploring even more of life's joys together with you!

I am immensely grateful to Jenny for her kindness and care during the final, intense stretch of writing. Thank you for ensuring I wouldn't starve, for offering sound advice, and for extending your endless compassion when it was most needed.

A special thank you goes to Tizoc for always supporting me and, especially in a clutch, taking over as a carry. May your ice machine never break, and may your bristles stay sharp.

To Wulf, thank you for the support and coffee offered. I am glad to count you to my friends and look forward to many more hours of coffee, camera, or 3D printing nerdiness.

On the subject of coffee, I also want to thank Faruk. Whenever I needed a change of scenery to reignite my motivation, Café Mundus provided the perfect atmosphere for creativity to flourish. Your support during my defense, from coffee processing discussions to roast sample preparations, was invaluable.

Of course, a big and heartfelt thank you goes to my parents, Artur and Monika, my grandmothers, Ruth and Irene, and my brother, Leonard. Your unwavering support, belief in me, and the environment you created allowed me to pursue my dreams. Thank you for always having my back, for your understanding, and for nurturing me into the person I am today.

Table of Contents

List of Figures

List of Tables

Abbreviations

AECC	**A**merican-**E**uropean **C**onsensus **C**onference
AI	**A**rtificial **I**ntelligence
AKI	**A**cute **K**idney **I**njury
ARDS	**A**cute **R**espiratory **D**istress **S**yndrome
ASIC	**A**lgorithmic **S**urveillance for **I**ntensive **C**are Units
AUC	**A**rea **U**nder the **C**urve
AUCPR	**A**rea **U**nder the **C**urve for **P**recision-**R**ecall
BGA	**B**lood **G**as **A**nalysis
BMI	**B**ody **M**ass **I**ndex
CART	**C**lassification **A**nd **R**egression **T**rees
CDSS	**C**linical **D**ecision **S**upport **S**ystem
CoI	**C**atalogue **o**f **I**tems
CSV	**C**omma **S**eparated **V**alues
CT	**C**omputed **T**omography
DDI	**D**rug-**D**rug **I**interaction
DEA	**D**iagnostic **E**xpert **A**dvisor
DT	**D**igital **T**win
EC	**E**thics **C**ommittee
ECMO	**E**xtra **C**orporeal **M**embrane **O**xygenation
EHR	**E**lectronic **H**ealth **R**ecord
FAIR	**F**indable **A**ccessible **I**nteroperable **R**eusable
FLOPS	**F**loating-**P**oint **O**perations **P**er **S**econd
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
GDPR	**G**eneral **D**ata **P**rotection **R**egulation

HFNO	**H**igh-**f**low cannula **n**asal **o**xygen therapy
HPC	**H**igh-**P**erformance **C**omputing
ICD	**I**nternational Statistical **C**lassification of **D**isease and Related Health Problems
ICU	**I**ntensive **C**are **U**nit
KDE	**K**ernel **D**ensity **E**stimation
LDA	**L**inear **D**iscriminant **A**nalysis
LIPS	**L**ung **I**njury **P**rediction **S**core
LOS	**L**ength **O**f **S**tay
LLM	**L**arge **L**anguage **M**odel
MII	**M**edical **I**nformatics **I**nitiative
ML	**M**achine **L**earning
MONAI	**M**edical **O**pen **N**etwork for **A**rtificial **I**ntelligence
MV	**M**echanical **V**entilation
NHR4CES	National High-Performance Computing for Computational Engineering Sciences
PEEP	**P**ositive **E**nd **E**xpiratory **P**ressure
ROC-AUC	**R**eceiver **O**perator **C**haracteristics **A**rea **U**nder the **C**urve
ROS	**R**andom **o**ver **S**ampling
RUS	**R**andom **U**nder **S**ampling
RWD	**R**eal **W**orld **D**ata
RWE	**R**eal **W**orld **E**vidence
RWTH	**R**heinisch-**W**estfaelische **T**echnische **H**ochschule
SDL	**S**imulatian and **D**ata **L**ab
SFTP	**S**ecure **F**ile **T**ransfer **P**rotocol
SMITH	**S**mart **M**edical **I**nformation **T**echnology for **H**ealthcare
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
UI	**U**ser **I**nterface
UKA	**U**niversitäts**k**linikum **A**achen (University Hospital Aachen)
VILI	**V**entilator **I**nduced **L**ung **I**njury
VP	**V**irtual **P**atient

Original Publications

Various publications resulting from the work af the Joint Research Center for Computational Biomedicine are referenced throughout this thesis. This list provides an overview of the publications and their use in the context of this thesis.

- Chapter 1 uses parts of the literature review from [7], [11], and [12] to introduce this work and the scientific context.
- Chapter 2 uses [7] to detail the structure of the ASIC project, under which most of the work presented here was developed.
- Chapter 2.1.1 references the thesis of Dr. Sharafutdinov [13] for details on data filtering and preprocessing. The filters were implemented colaboratively in the context of the ASIC research project. Tables and Figures are recreated here with permission. More detail is provided in his work.
- Chapter 3 uses parts of [12]
- Chapter 4 uses parts of [11]

[7] Marx, G., Bickenbach, J., Fritsch, S.J., Kunze, J.B., Maassen, O., Deffge, S., Kistermann, J., Haferkamp, S., Lutz, I., Voellm, N.K., Lowitsch, V., Polzin, R., Sharafutdinov, K., Mayer, H., Kuepfer, L., Burghaus, R., Schmitt, W., Lippert, J., Riedel, M., Barakat, C., Stollenwerk, A., Fonck, S., Putensen, C., Zenker, S., Erdfelder, F., Grigutsch, D., Kram, R., Beyer, S., Kampe, K., Gewehr, J.E., Salman, F., Juers, P., Kluge, S., Tiller, D., Wisotzki, E., Gross, S., Homeister, L., Bloos, F., Scherag, A., Ammon, D., Mueller, S., Palm, J., Simon, P., Jahn, N., Loeffler, M., Wendt, T., Schuerholz, T., Groeber, P., Schuppert, A., 2021. Algorithmic surveillance of ICU patients with acute respiratory distress syndrome (ASIC): protocol for a

multicentre stepped-wedge cluster randomised quality improvement strategy. BMJ Open 11, e045589. https://doi.org/10.1136/bmjopen-2020-045589

[11] Polzin, R., Fritsch, S., Sharafutdinov, K., Bickenbach, J., Marx, G., Schuppert, A., n.d. Predicting a sudden decrease in oxygenation in mechanically ventilated intensive care patients as a surrogate marker for acute respiratory distress syndrome. [*Unpublished Manuscript*].

[12] Polzin, R., Fritsch, S., Sharafutdinov, K., Marx, G., Schuppert, A., 2023. Diagnostic Expert Advisor: A platform for developing machine learning models on medical time-series data. SoftwareX 23. https://doi.org/10.1016/j.softx.2023.101517

[14] Sharafutdinov, K., Bhat, J., Fritsch, S., Nikulina, K., Samadi, M., Polzin, R., Mayer, H., Marx, G., Bickenbach, J., Schuppert, A., 2022. Application of convex hull analysis for the evaluation of data heterogeneity between patient populations of different origin and implications of hospital bias in downstream machine-learning-based data processing: A comparison of 4 critical-care patient datasets. Frontiers in Big Data 5, 603429. https://doi.org/10.3389/fdata.2022.603429

[15] Sharafutdinov, K., Fritsch, S., Iravani, M., Farhadi, P., Saffaran, S., Bates, D., Hardman, J., Polzin, R., Mayer, H., Marx, G., Bickenbach, J., Schuppert, A., 2023. Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets. IEEE Open Journal of Engineering in Medicine and Biology PP, 1–11. https://doi.org/10.1109/OJEMB.2023.3243190

[16] Meyer, J., Fritsch, S., Sharafutdinov, K., Nikulina, K., Polzin, R., Marx, G., Bickenbach, J., Schuppert, A., 2023. Transfer Learning Across Diseases Opens a Novel Route Towards Pandemic Preparedness [*Preprint*]. https://doi.org/10.21203/rs.3.rs-3349295/v1

[17] Linden, T., Ku, C., Wendland, K., Sharafutdinov, K., Polzin, R., Schuppert, A., Fröhlich, H., n.d. Survival Multi-Modal Neural Ordinary Differential Equations for Mortality Prediction of Patients with Severe Lung Disease [*Unpublished Manuscript*].

Chapter 1

Introduction and Background

Acute Respiratory Distress Syndrome (ARDS) is a life-threatening condition that affects millions of people worldwide, resulting in significant morbidity and mortality [1]. Despite advances in medical technology and treatment strategies, the mortality rate for ARDS remains high, with up to 46% of patients dying [2]. One of the major challenges in treating ARDS is identifying patients at high risk of developing severe respiratory failure, which could help clinicians intervene early and potentially prevent or mitigate the progression of the disease. ARDS is an example for a wide class of life-threatening syndromes also including, for example, sepsis. For these syndromes, survival decreases rapidly with time to therapeutic intervention. They manifest within short time frames out of apparently 'thin air', but in through early interventions severe disease courses can often be avoided. Hence, early prediction is crucial for improving therapeutic strategies in critical care.

Currently, the diagnosis of ARDS relies on clinical evaluation and radiological findings, which can be time-consuming and subjective [18]. There is a pressing need for a more accurate and objective method to predict the risk of developing ARDS and loss of oxygenation in very early states to start lung-protective measures. Diagnostic prediction models assess the current health status of a patient [19] [20]. This work presents such a model, which, in contrast to diagnosing ARDS when it is already present, incorporates various physiological and demographic factors, to identify patients at high risk of developing ARDS. Thus, the new model enables targeted interventions to prevent or mitigate the progression of the disease.

While diagnosis is of crucial importance to inform subsequent healthcare decisions and explain a patient's health problems [21], predicting upcoming health situations allows medical experts to act early on and focus their attention where needed. Timely recognition of such events can prevent catastrophic events in the ICU [22].

This section presents related work and places this thesis into context. It first provides an introduction to AI, in general, in healthcare, and as a driver for alarm systems in the ICU, such as the system developed in Chapter 4. The next part of this section details the syndrome that motivates the development of the model: acute respiratory distress. Its impact on global health is outlined, a short medical background is provided, and the current definition of ARDS is detailed. Further, considerations and complications in predicting the onset of the syndrome are discussed. Finally, this section provides context for the platform for medical timeseries research presented in Chapter 3, by highlighting the specific needs it satisfies, and the gap in technology it fills. The importance of high-performance computing (HPC) in ML research, especially in a clinical setting, is discussed, and the need for research software is emphasized.

1.1 Artificial Intelligence in Intensive Care

The increase in computational capacity over recent years allowed artificial intelligence (AI) to enable leaps in research in various disciplines. From self-driving cars [23] to video synthesis from text input [24], more and more astonishing breakthroughs are happening every year. AI is foretold to transform science [25], and with around 8% of research papers mentioning AI or machine learning (ML) terms in their title or abstract, AI is already impacting a significant share of today's science landscape [26]. The term AI often refers to computers or other digital devices that can operate in a way that seems intelligent to us, for example, reading, writing, talking, playing, or making recommendations [27]. Among many other areas, the broad field of AI encompasses the area of machine learning, which is a field of study that is focused on learning from data to perform tasks on previously unseen data. ML algorithms are nowadays being used in many fields, such as computer vision, e-commerce, marketing, natural language processing, cybersecurity, finance, and healthcare.

Artificial intelligence in medical and healthcare applications is growing rapidly and there is a substantial concentration of global funding in this sector [28]. In the health care field AI is being increasingly used [29]. For example, in diagnostic imaging, medical experts could be assisted in the evaluation of X-ray images to improve detection rates or highlight areas of concern [30].

Drug discovery and development is accelerated through AI models analyzing large databases or predicting the biological effects of new drugs [31]. AI could even help communication between patients and medical doctors, with more empathic and elaborate replies to questions [31]. Other applications range from surgery assistance through augmented reality [32], or robotic surgery [33], to mental health applications [34], and overall healthcare operations management [35].

In the context of this work, we focus on the application of ML methods for predictive analytics for patient outcomes in critical care. This area of research is concerned with the analysis of patient data to predict disease progressions and outcomes by monitoring heterogeneous data, which are assessed with highly different sampling rates throughout the hospital stay. This allows for an early identification of high-risk patients and corresponding reactions, such as proactive treatment, close monitoring, or the enlistment into research studies.

While the deteriorating states of patients in general hospital wards often go unnoticed for prolonged periods of time [36], patients in ICU wards are continuously monitored, generating large amounts of data with thousands of measurements per patient per day. Predictive analysis be used as a way to digest and analyze this data and predict various endpoints to improve patient outcomes, such as upcoming septic shock or ARDS. Proactively directing the medical expert's attention to patients with a high risk of upcoming complications can allow for a treatment at earlier, often less severe, stages of disease progression, or even prevent the rise of complications altogether.

Data for training of ML models can be collected either through clinical trials, the traditional route, or by retrospective observational data from the "real world", so-called real world data (RWD). While traditional clinical trials often collect data in dedicated settings with a carefully chosen population and special attention to capturing the relevant parameters, RWD in contrast, snapshots the clinical reality [37]. Clinical evidence derived from analyzing RWD is referred to as real-world evidence (RWE) [38].

One approach to integrating predictive modeling into the ICU environment is developing clinical decision support systems (CDSS) to provide information to doctors on the bedside. Computerized DSS have been utilized in clinics since the 1980s and are now widely used. In 2013, CDSSs were present in roughly 40% of American hospitals [39]. While the scope of CDSS can vary widely, ranging from diagnostics and disease management to computerized guidelines for documentation [40], this work is mostly concerned with CDSS providing an alarm system. CDSS alarm systems can significantly improve patient care.

One example is in the context of drug-drug interaction (DDI), where errors are often cited as common and preventable. Up to 65% of inpatients have historically been exposed to potentially harmful combinations of drugs [39]. Implementing CDSS to safeguard dosing, duplication, and DDI significantly impacts the outcome, improving prescribing, reducing side effects and drug interactions, and improving the overall patient outcome [41].

While, historically, CDSS were often implemented as bedside monitors, this is nowadays shifting to web applications and integrations with electronic health records (EHR), which can be accessed through a variety of devices, ranging from smartphones or tablets to desktop computers, biometric monitors or wearable health technology [42]. CDSS can be roughly categorized as data-, or knowledge-based. Knowledge-based systems are designed to follow expert medical knowledge, often implemented in the form of if-then rules [43]. On the other hand, data-based systems digest a data source to produce outputs based on patterns recognized in the data. They often utilize AI, ML, or statistics to generate their decision support and pose various challenges that are not present in knowledge-based systems. Lack of explainability in such black-box models, which are quite common in complex and deep machine learning models, is considered by some to be fundamentally violating the principles of medical ethics [44]. Looking at existing bias in available training data, AI models trained on this data can learn to reproduce it, for example, in biases against Black patients [45]. Aside from the challenges arising from black-box systems in data-based systems, there are various more generic pitfalls when it comes to CDSS. They can disrupt workflows [46], impact user skill [47], raise further interoperability problems [39], and the overall financial viability sometimes remains a struggle [48]. One additional issue with CDSS is often referred to as alarm fatigue.

Alarm systems in the ICU have been used for a long time, with simple versions sounding an alarm whenever patients' vital parameters significantly deviate from the normal range. Over time, more and more systems have accumulated, which will raise alarms in specific situations. Examples of equipment often integrating such CDSS include pulse oximeters, catheters, mechanical ventilators, and electrocardiograms [49]. Up to 40 diverse alarms of different monitoring systems may be present in an ICU [50]. These alarms are often tuned to be highly sensitive to reduce the risk of missing life-threatening events. But this will, in turn, reduce the specificity and increase the rate of false alarms [51]. The result is a high level of noise and false alarm rates of up to 88% [52]. With volumes regularly exceeding 80dB [53], and often more than a hundred alarms per bed per day [54], the clinicians can over time desensitize to the alarms and the stress of the patients increases [55] [56] [57] [58] [59]. This is commonly referred to as alarm fatigue.

False alarms in the ICU are considered a top health technology concern, with the ECRI Institute listing alarms as one of the top 10 Health Technology Hazards every year since the list has first been published in 2007 [60] [61].

Considering the dangers of alarm fatigue, it is necessary to judge models intended for use in the ICU especially carefully, balancing the risk for patients in critical states without alarm on the one side and alarm fatigue on the other side. In the already noisy environment of the ICU, a high level of model performance is especially desirable. The risk of missing an event of interest, in contrast to the issues caused by alarm fatigue, is highly individual and depends on the patient's current state. For a patient that is in a less dangerous situation, it might be suitable to lower the sensitivity of the model, reducing the rate of false alarms in the process.

1.2 Acute Respiratory Distress Syndrome

ARDS was first described as a "wet lung" or "shock lung" in 1821 by Laennec et al. [62]. The term ARDS was coined in 1967 by Ashbaugh et al. and defined as "acute onset of tachypnea, hypoxemia and loss of compliance after a variety of stimuli" [63]. In 1994, the American-European Consensus Conference Committee (AECC) recommended a clearer definition of ARDS [64], which remained state of the art until 2012. Various shortcomings, such as inconsistencies in severity grading or a lack of definition for clear onset times, were addressed in the Berlin Definition in 2012 [65]. The Berlin Definition categorizes three different severities of ARDS and describes the requirements for a corresponding diagnosis. Most notably, an ARDS diagnosis based on the 2012 Berlin definition requires the following:

- An acute onset lung injury within one week of an apparent clinical insult and with the progression of respiratory symptoms
- Bilateral opacities on chest imaging are not explained by other lung pathology.
- Respiratory failure is not explained by heart failure or volume overload.
- A decreased Horowitz index (PaO_2/FiO_2 ratio)
  - Mild ARDS: 201 - 300 mmHg (39.9 kPa)
  - Moderate ARDS: 101 - 200 mmHg (26.6 kPa)
  - Severe ARDS: 100mmHg (13.3 kPa)
- All while a minimum positive end-expiratory pressure (PEEP) of at least five cmH_2O was present.

The Horowitz index is defined as the ratio of partial pressure arterial oxygen (PaO_2)

and the fraction of inspired oxygen (FiO_2) [66] [67]. It compares blood oxygen levels and the oxygen concentration in the breath. Arterial blood is analyzed during a blood gas analysis (BGA) to measure PaO_2. FiO_2 is defined by the level of oxygen enrichment of air a ventilated patient is breathing. While the Berlin definition improves on the AECC definition of ARDS, there are still many areas that need improvement. One example is that the severity levels defined in the Berlin definition seem to fail to assess the severity of lung injury [5] [68]. Another example stems from a change in the reality of clinical respiratory care. High-flow nasal cannula oxygen therapy (HFNO) has been used for managing severe acute hypoxemic respiratory failure more frequently while achieving PEEP and FiO_2 levels that are not well fit for the criteria defined in the Berlin definition [69]. Finally, the overall health of the lung decreases with age and older patients might have an overall lower Horowitz index. This is not well represented in the Berlin definition.

Acute respiratory distress syndrome is a severe form of respiratory failure, associated with high mortality in patients in the ICU ranging from 25% to 46% [2]. ARDS is often under diagnosed [70] and the reported incidences and mortality vary significantly [71] [72] [73] [74] [75] [76]. The prevalence of ARDS is high, with up to three million cases per year and 10% of all ICU admissions showing ARDS at some point during their time in hospital [1] [2]. It is considered one of the most common reasons for admission to the ICU [77].

ARDS is a widespread inflammation of the lung that impairs its ability to exchange oxygen and carbon dioxide. The syndrome results from lung injury with various causes, such as sepsis, pneumonia, trauma, aspiration, or pancreatitis. Hypoxemia, an oxygenation deficiency in the blood [78], is often found in ARDS patients, as well as atelectasis, where pulmonary alveoli, responsible for gas exchange in the lungs, partially collapse.

The primary treatment involves mechanical ventilation (MV) and adjunctive therapies, such as prone positioning, as well as treating the cause of the syndrome [79]. Neuromuscular blockers and lung recruitment maneuvers are also used as treatment strategies. If these treatments still do not sufficiently increase the oxygenation, extracorporeal membrane oxygenation (ECMO) represents a very invasive treatment option [1] that can be used to further increase oxygenation. ECMO is a type of artificial life support, where a machine takes over functions of the heart and lung.

Mechanical ventilation, while often a life-saving intervention, bears the risk of patients developing ventilator-induced lung injury (VILI) during treatment. Lung protective ventilation has become a standard treatment strategy, in part to minimize or prevent additional damage to

the lungs during ventilation. However, adherence is still low [2]. Early detection is considered essential to prevent a progression of deterioration [3]. ARDS is present in almost a fourth of all mechanically ventilated patients. However, it is often under diagnosed and frequently detected delayed [2].

Research has shown that delayed treatment of patients with ARDS has a significant impact on the associated mortality [4]. Few effective therapeutic modalities exist to deal with ARDS [80]. Treatment thus is often focused on preventing additional lung damage, which accumulates rapidly if left untreated [81]. Therefore, detecting ARDS reliably and during an early stage has a major impact on the further development of a patient's ARDS, with the first days after onset representing an important therapeutic window [5] [6]. Hence, early prediction of upcoming ARDS is crucial for ventilation management, with a high potential impact on the overall outcome. With no effective pharmacologic treatments targeting the underlying pathology, lung-protective management of ventilation currently presents the best strategy.

1.3 Developing ML models on medical timeseries data

Machine learning in the context of healthcare data, especially in areas like intensive care, where large volumes of data are generated, requires extensive computational resources. Driven by the size of the datasets and the complexity of deployed algorithms, many ML models stand to benefit from HPC, with some areas, such as deep learning, often outright requiring HPC.

Broadly speaking HPC refers to the application of computers with high levels of performance, nowadays often clustered and used in parallel, to solving computationally demanding problems. The performance of a computer can be measured in floating-point operations per second (FLOPS). While personal computers achieve around 10^{11} to 10^{13} FLOPS [82], supercomputers can achieve more than 10^{18} FLOPS [83]. Many computationally expensive tasks and research areas rely on HPC, as personal computers quickly hit their limits with regard to storage, throughput, memory, or raw computational power. Training of large language models, for example, can take up to 3,640 Petaflops/s-days [1] [84], requiring hundreds to thousands of GPUs working in parallel [85]. HPC facilities can provide the necessary resources to support these efforts.

Common areas of research that utilize HPC regularly include, for example, combustion or fluid simulation [86], material design [87], geological disaster recognition [88], or protein struc-

[1] Each Petaflop/s day is 8.64×10^{19} flops.

ture prediction [89]. Machine learning is nowadays being used widely throughout the healthcare field. Applications range from diagnostic imaging, where medical experts are assisted in evaluating X-ray images [90] [91], to Surgery assistance, where augmented reality overlays information during the procedure [92]. With large language models (LLM) becoming increasingly prominent, even more areas of healthcare are beginning to utilize AI. For example, the communication between patients and medical doctors can be improved, with ML helping in the creation of more empathic and elaborate replies [93], or mental health applications, such as stress-, depression-, or suicidality-detection [94]. The widespreach clinical adoption faces various challenges though, for example the need for specialized infrastructure and expertise [95].

In the context of this work, we focus on the application of ML methods to medical timeseries data, an area of research with significant interest of the worldwide community [96] [97] [98] [99] [100]. A lot of the research work is focused on the development of dedicated methods with a focus on solving specific tasks in specific settings. For example, the prediction of upcoming Sepsis, ARDS, or Acute Kidney Injury (AKI). Research is also oriented along different steps of the research pipeline, such as work on patient data processing [101] [102] or parameter imputation [103] [104]. Jarrett et al. released a pipeline toolkit for medical timeseries that presents "the first comprehensive and automatable pipeline for clinical timeseries ML"[105]. This software aims to provide an integrative end-to-end solution focusing on breadth and flexibility, covering as many steps as possible, from preprocessing and feature selection to model calibration and various prediction targets. Other existing solutions usually focus on specific subtasks, such as imputation and prediction [106] [107] [108] [109] [110] [111].

While various frameworks for working with timeseries data exist, to our knowledge none of them directly integrate a simplified interaction with HPC systems and provide a flexible and interactive visualization component. The need for such software arises at the intersection of ML research and clinical expertise, with researchers and physicians cooperating in joint projects. Chapter 3 of this work presents the Diagnostic Expert Advisor, a platform that fulfills this need by focusing on the ease of HPC use, a patient-centric workflow, and deeply integrated visualizations. In contrast to the existing solutions, the proposed platform neither targets only specific subtasks, nor attempts to provide an end-to-end solution. It instead offers a template to quickly base research work on, as well as a flexible standard to integrate individual methods for various specific tasks. The proposed platform could, for example, integrate models for Sepsis, ARDS, and AKI, using all the abovementioned tools for processing and imputation, would ensure a common data format and patient-centric workflow is utilized, and encourage extensive visualizations and the

integration of HPC. It has been used to develop a prediction model for ARDS in intensive care as detailed in Chapter 4.

Nowadays, software is at the core of many research projects, constituting a significant pillar of the academic environment in various domains. Such software is an outcome of research work that is regularly neglected in contrast to the papers presenting the results [112]. There is a clear trend, with research finding the percentage of papers including code at some major conferences doubling over six years, and a clear impact of code release on citation numbers [113] [114]. Yet many publications still do not include any code at all. A potential reason for this is lack of time or insufficient training [115] [116]. One approach to improve on the current state is by establishing platforms that encourage good coding practices, such as separation of concerns and modularization.

1.4 Outlook

This work first establishes a research platform to provide a fast and efficient way for researchers to develop machine learning models on heterogeneous time series data in medicine, especially for rapidly developing and life-threatening syndroms and conditions. Then such a prediction model is presented, focusing on rapid loss of oxygenation in mechanically ventilated intensive care patients. The proposed prediction model enables clinicians to identify high-risk patients early and provide personalized care and treatment.

This chapter outlined the motivation for and contributions of this thesis. It further provided the necessary background to place the work into context. AI, in general, and in the context of medical research, with a focus on CDSS, was discussed. The relevance and definition of HPC was briefly presented, and the special environment of the ICU was detailed. ARDS was introduced, and the relevance of research software was highlighted.

Chapter 2 describes the data used throughout this work and provides an overview of the patient cohorts defined. It further elaborates on the processing of this data and its utilization in modeling.

Chapter 3 presents a software platform for the development of ML models on timeseries data in medicine. This platform enables researchers to utilize HPC for development more easily and encourages an interdisciplinary workflow with a focus on intuitive data handling and visualization.

Chapter 4 elaborates on the development of a ML model to predict a decrease in the lungs capability to oxygenate the blood in intensive care patients. This model predicts a derived surrogate marker that can indicate the onset or worsening of ARDS in patients. The model is evaluated on the cohort introduced in Chapter 2, and its retrospective feasibility in the context of the COVID pandemic is tested by stratifying the cohort with respect to the presence of SARS-CoV-2.

Chapter 5 discusses the overall results and this works contribution to the research landscape. It highlights and summarizes the implications of both, the developed software platform and the predictive CDSS model.

Chapter 2

Data

Healthcare institutions generate large amounts of heterogeneous data [117] which could reveal health patterns and provide novel solutions to a diverse range of problems [118]. Accessing and conducting research using this data is often not trivial. Privacy and security requirements complicate access, and data ownership questions are often not fully clarified. The data used in this research originates from two different sources, spanning three individual datasets, which will be explained in more detail in this chapter. Most notable Chapter 2.1 details the "Algorithmic Surveillance for Intensive Care Units" (ASIC) study, in whose context most of this research took place. The processing and preparation of this data is described in Chapter 2.1.1. Specific adaptations for the prediction model are later detailed in Chapter 4. Chapter 2.1.2 then touches on potential biases when working with data from different hospitals. The freely available MIMIC-III dataset is detailed in Chapter 2.2. This section describes its uses in the context of this work and highlights differences in contrast to the previously introduced ASIC dataset. Chapter 2.3 details challenges faced when working with imbalanced data, and finally, Chapter 2.4 summarizes this chapter.

2.1 Algorithmic Surveillance for Intensive Care Units

In February 2021 Marx et al. published the protocol for the "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome" (ASIC) study [7] as part of the "Smart Medical Information Technology for Healthcare" (SMITH) consortium which is part of the "Medical Informatics Initiative" (MII) of the German Federal Ministry of Education and Research.

The protocol describes the use of a mobile app to support the timely detection and treatment of ARDS in ICU patients, not unlike a knowledge-based CDSS for ARDS. Overall, eight German university hospitals participated in the project, with 31 ICUs in total. Data for the project was collected over multiple months, with an 18-month preparatory phase, an eight-month control phase, and a three-month roll-in phase, followed by a quality improvement phase. An overview of the participating hospitals and the respective ICUs and beds included in the study can be seen in Table 2.1. Patients who were 18 years or older were included in the study if they were mechanically ventilated for at least 24 hours. Data collected for these patients includes demographic data, such as age, sex, or weight, routinely charted ICU variables, such as heart rate or body temperature, and International Statistical Classification of Disease and Related Health Problems (ICD)-10 codes. An overview of the parameters used from this study is available in the Appendix. Both, the independent Ethics Committee (EC) at the RWTH Aachen Faculty of Medicine (local EC reference number: EK 102/19) and the respective data protection officer, approved the ASIC project in March 2019. It is registered at the German Clinical Trials Register (DRKS00014330) and was conducted according to the Declaration of Helsinki [119]. As the app merely supplements the existing EHR, the need for informed consent was waived for the routine data collected in this study [7].

The project was designed as a multicentre stepped wedge cluster randomized trial [120]. In total, 12 clusters were created, with ICUs that are dependent, either technically or organizationally, summarized into clusters. An initial control phase captures the state of care. It is followed by a roll-in phase, where the ASIC app is implemented. For this roll-in phase, the randomized clusters are added in a stepwise fashion, with a one-month delay between every addition.

It is notable, that the Coronavirus disease 2019 (COVID-19) pandemic impacted the study, with data collection scheduled from July 2019 to December 2021, with the first COVID-19 patient in Germany being confirmed on 27 January 2020 near Munich, Bavaria [121]. COVID-19 is generally being associated with ARDS [122], thus having a significant impact on various aspects of the ASIC dataset.

Table 2.1: Overview of the hospitals participating in the ASIC studies and the respective count of ICU wards and total beds.

Hospital	ICUs	Beds
Aachen	8	96

Hospital	ICUs	Beds
Bonn	4	58
Düsseldorf	1	16
Halle a.d. Saale	2	30
Hamburg-Eppendorf	10	116
Jena	2	50
Leipzig	3	78
Rostock	1	23

The patient data collected in the ASIC study was pseudonymized to protect personal information. Different definitions exist with respect to anonymization and pseudonymization. The EU General Data Protection Regulation (GDPR) defines anonymous information as "[…] information which does not relate to an identified or identifiable natural person […]." [123]. In contrast to anonymized data, pseudonymization is defined as "[…] the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information" [124]. To summarize, pseudonymization leaves personal information in a state where no individual can be identified without additional information. This additional information is very relevant to the distinction between pseudonymization and anonymization. Anonymization is, in the context of GDPR, an irreversible process, preventing the identification of individuals at the current state of technology. On the other hand pseudonymization allows for the re-identification of individuals if additional information is available. In the context of clinical research an example for such additional information could be a table containing a mapping of in-hospital patient identifiers and pseudonymized identifiers. Re-identification would thus be possible, but not without the additional information.

Preparing medical data in a way that identification is not possible is a complex task. Considering the pseudonymization of neurological images for example, the accepted state of the art for facial feature removal provided adequate pseudonymization to publicly share brain scans. Through the advances in machine learning, and deep learning in particular, the possibility of re-identification of such images is increasing though [125]. Some data may not event be possible to anonymize fully. It is, for example, still up for debate, whether it is possible to fully anonymize brain images in general. There is a clear consent though, that high levels of anonymization might hinder the scientific usability of data as more and more information is removed.

In the context of the ASIC project it was ensured that the data shared possesses the property of *k-anonymity* [126]. This property is achieved if, in a given dataset, it is not possible to distinguish each person from at least $k-1$ other persons in the data. This leads to many of the demographic data being clustered, such that a combination of e.g. age, weight, sex and height can not be used to re-identify a patient. Table 2.2 provides an example of how patient data is altered during such a pseudonymization process. It also highlights different issues arising with k-anonymity. Most notably, it is both crucial and complicated to find all parameters that can be used to identify patients. While e.g. names and addresses are quite obviously identifying, parameters such as age, height, and weight can be *quasi-identifiers*. These need to be processed to ensure every distinct set of quasi-identifiers is not unique, but possesses the property of k-anonymity. If any quasi-identifiers are missed, there is a risk of de-identifiability [127] [128].

Table 2.2: Excerpt example of pseudonymized data with k-anonymity

Parameter	Description	Originaly reported	Pseudonymized
Weight	Weight of the patient	70kg	50-120kg
BMI	Body-Mass-Index	18.5	Normal
Age	Age of the patient	74	50-80
Height	Height of the patient	193	>190cm

Table 2.3: Example of the ASIC data format.

Time from admission (min)	FiO2 (%)	PaO2 (mmHg)	Heartrate (bpm)	…	Body Temperature (°C)	Lipase (U/L)
0	80	70	101	…	37	None
15	83	None	104	…	37	None
30	87	None	97	…	37	None
45	85	None	100	…	38	110
60	80	70	105	…	38	None
75	72	None	108	…	38	None
90	70	None	111	…	39	None
105	60	None	132	…	39	40
120	55	120	127	…	39	None
135	52	200	140	…	39	None
…	…	…	…	…	…	…

The data gathered from the ASIC study is available in comma-separated values (*csv*) format, with every row denoting a 15-minute interval. All data was collected by the

Data Integration Center of the University Hospital RWTH Aachen and transferred to the Joint Research Center for Computational Biomedicine via a SFTP server. An example of the format can be seen in Table 2.3. While the full ASIC cohort, denoted as the ASIC *Control* data, was only available at the end of the study, a dedicated *Calibration* cohort was created, containing retrospective data from each participating hospital. The intended use of the *Calibration* dataset was foremost the identification and evaluation of any biases that could impact the final *Control* dataset. For this calibration dataset a minimum of 1,000 patients, meeting the inclusion criteria ($\geq$ 18 years and $\geq$ 24h cumulative mechanical ventilation), for each hospital were selected. Sharafutdinov et al. worked extensively with the *Calibration* cohort during the development of a pipeline for generalization assessment [13] [14].

This work focuses on the development of a prediction model for rapid loss of oxygenation. For two reasons this development utilizes only data from a single hospital. On the one hand, due to delays in the data collection and quality assurance process, the quality of data available for different locations varied widely. On the other hand, various issues can arise by pooling of data from different hospitals, as further detailed in Chapter 2.1.2.

Thus, the *Control* data from the University Hospital Aachen (UKA) was used, providing the largest and most complete contribution to the ASIC cohort at the time of writing. Further, cooperation within the ASIC project included medical expert from this location, enabling fast and direct collaboration. An overview of the population of the data collected for this hospital can be seen in Table 2.4. An overview of the data from all hospitals in the ASIC study, at the time of writing, can be found in the Appendix.

Table 2.4: ASIC University Hospital Aachen Data Overview.

Total Encounters, n	3,676
Male Gender, n (%)	65
Length of ICU Stay, days (mean±std)	20.5 ± 21.8
Mortality, n (%)	1,220 (44)

2.1.1 Data Preprocessing

Real-world data brings many challenges, such as the variability in data quality [129]. We thus implemented filters for the ASIC data to remove unphysiological values for many of the parameters. Most of the variables were thresholded based on discussions with medical experts, but a more sophisticated algorithm was developed for the Horowitz index, which is of crucial importance for the proper detection of ARDS. Most of the work on the ASIC data was conducted in close cooperation with my dear colleague Dr. Sharafutdinov, who elaborates on the Data Filtering in great detail in Chapter 3.3 of his thesis [13]. The thresholds utilized are listed in Table 2.5.

Table 2.5: Thresholds for parameters in ASIC preprocessing. Recreated with permission ([13]).

Paramater	Lower threshold	Upper threshold
Central venous oxygen saturation (ScvO2), %	40.0	90.0
Haemoglobin (Hb), mmol/L	2.0	10.0
FiO2, %	20.0	100.0
Horowitz index, mmHg	10.0	1500.0
PEEP, cmH2O	2.5	30.0
PEI, cmH2O	5.0	45.0
Respiratory rate, 1/min	5.0	40.0
Body temperature, C	30.0	45.0
Tidal volume, ml	100.0	1000.0
Base excess (arterial), mmol/l	-25.0	25.0
SaO2, %	85.0	100.0
pH (arterial), unitless	6.0	8.0
Bicarbonate (arterial), mmol/l	10.0	50.0
paCO2, mmHg	15.0	90.0
paO2, mmHg	50.0	250.0
Inspiration : Expiration ratio (I:E), unitless	0.3	6.0
Cardiac output, l/min	1.5	20.0
Driving pressure (deltaP), cmH2O	2.5	100.0

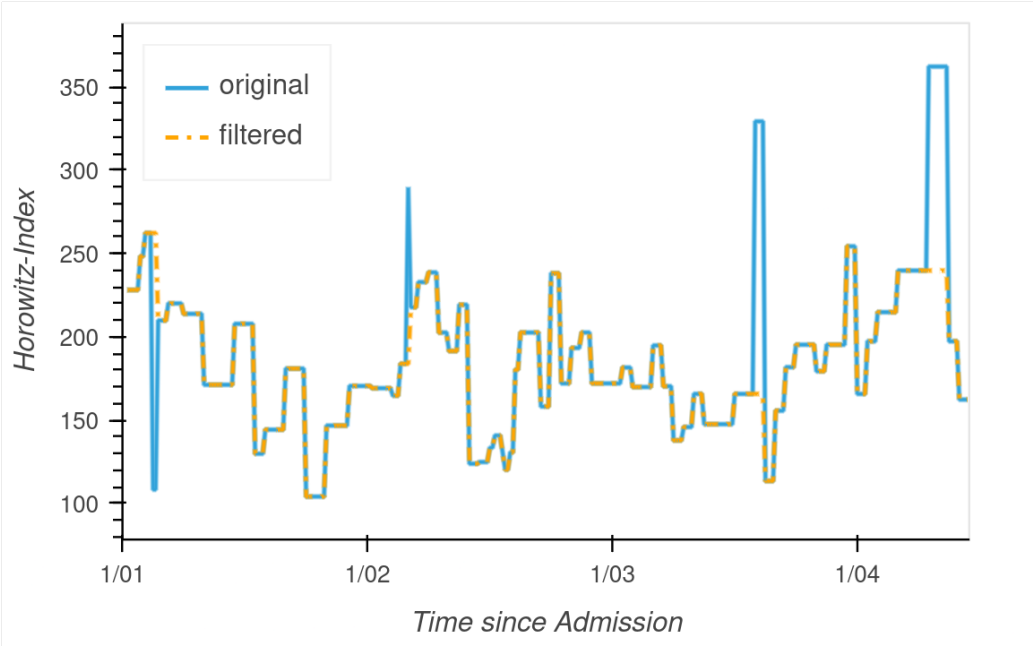


Figure 2.1: Example of Horowitz index filtering. The blue line denotes the original data, including unphysiological jumps in Horowitz index, which are removed by the filtering.

With respect to Horowitz index most notably an algorithm for the retrospective calculation of Horowitz index if PaO_2 and FiO_2 were present was developed. If the Horowitz index was missing in the original data it was calculated. Calculation used the nearest FiO_2 value to every PaO_2, if it was no older than eight hours. This algorithm was developed in close cooperation with medical experts in the ASIC workgroup. Listing 2.1 outlines the approach to retrospective Horowitz index calculation that was deployed. We further implemented a filtering on the Horowitz index that replaced consecutive measures which differed by more than 100 mmHg with missing values to handle extreme outliers and unreasonable jumps. This filtering replaced 1% of Horowitz index measurements. An example can be seen in Figure 2.1.

Listing 2.1 Pseudocode for the retrospective calculation of the Horowitz index

```
for pi in EHR[PaO2]:
    if Horowitz not in EHR[pi.time]:
      fi = find_closest_FiO2(EHR, pi)
      if fi.time.difference(pi.time) <= 8 hours:
        EHR[pi.time].Horowitz = pi/fi
```

2.1.2 Biases across hospitals

When applying AI in healthcare across hospitals several risks and challenges arise. Some of which can severely impact the performance of a model and make it potentially even dangerous to use, if not properly validated in the target context. Hospitals may, for example, differ in patient demographics or healthcare practices. Biases towards a certain demographic or group which is much more or less present in one hospital may impact the performance of that group in another hospital significantly. Healthcare practices are often different between hospitals, with different treatment protocols, diagnostic criteria or routines. A model trained on one hospital may pick up patterns that can only be found in the specific environment of that one hospital and may perform significantly worse on hospitals employing other guidelines, different criteria, or following other routines. In addition to the aforementioned risks, different medical devices, operating with different precision, reporting frequencies, or applied differently, could also cause a significant change in the underlying distribution of some parameters, and could reduce the performance of a model. Yet another intuitive example of risks when applying models between hospitals is in the prevalence of diseases. A hospital located in a particular cold region could be treating significantly more patients for frostbite or hypothermia. These challenges should be considered, if a model is to be applied on hospital different from the one it was trained on. Sharafutdinov et al. focussed on techniques to analyze and mitigate such risks [14]. In the context of this work, the risks have been limited by focussing on a single hospital.

2.2 MIMIC-III

The MIMIC-III database [130] is one of the largest single-center ICU databases, containing 58,976 admissions of 46,520 patients in total. Demographics, vital parameters, lab tests, medications, as well as imaging reports and caregiver notes are included in the dataset. In between 2001 and 2012 the data was collected in the critical care units of the Beth Isreal Deaconess Medical Center in Bosten. All data is deidentified and made available to researchers under a data use agreement. Physiological data from bedside monitors is recorded hourly and International Classification of Diseases, Ninth Edition (ICD-9) Codes are provided. As of February 2024 the MIMIC-III publication has been

cited more than 3,000 times since it was published in 2016. An even more exhaustive database has since been released in the form of MIMIC-IV, containing patients from 2008 to 2019 and additional information [131].

After meeting all requirements to use MIMIC-III, the data was downloaded and hosted as a MySQL database with the IT-Center Aachen. A cohort of patients which fit the ASIC inclusion criteria ($\geq$ 18 years old and $\geq$ 24h cumulative mechanical ventilation) was selected as an initial dataset to work with, while data from the ASIC study was being collected. The parameters extracted from MIMIC-III were limited to ones similar to what was available in the ASIC cohort, any other information, such as textual reports or images, was not extracted.

Table 2.6: Statistics on the extracted MIMIC-III cohort and the Aachen University Hospital in the ASIC study. The average age can not be calculated for ASIC due to the clustering required for k-anonymity.

Datasource	Total Encounters	Average Age	Female %	Average LOS	Mortality
ASIC	3,676	-	35%	20.5 days	44%
MIMIC	7,683	64 years	43%	13.5 days	17%

The final MIMIC cohort contained 7,683 encounters in total. More details are shown in Table 2.6. Most notably the cohort collected for the ASIC study has a longer length of stay and mortality, with significantly more male patients.

In the context of this work the MIMIC-III database was used to develop the pipeline for data processing, set up model structures, and to develop the initial versions of the Diagnostic Expert Advisor [12]. Many aspects of the data format are similar between the MIMIC-III and ASIC data when developing ML models. Both cohorts contain on the one hand timeseries data, and on the other hand demographics, and were stored in a similar format. Thus, it was possible to set up and test different pipelines and model configurations.

An overview of the differences between the MIMIC and ASIC cohort used in this work is given in Table 2.7. With respect to model development and data handling, the most significant differences are in the resolution of the data and the version of ICD-

Codes used. The formating and naming of parameters also differs between the two cohorts. Hospitals in the ASIC study agreed on a well-defined common set of parameters, described in a Catalog of Items (CoI). With each individual hospital intended to internally then transform the data they provided to the desired format before providing it. The data model in the MIMIC cohort is structured to "balance simplicity of interpretation against closeness to ground truth" [130].

Table 2.7: Comparison of MIMIC and ASIC cohorts

	MIMIC-III	ASIC
Resolution	hourly	15-minutes
Trial Type	single center	multi center
Location	Boston, USA	multiple, Germany
Admissions	58,976	3,676 (UKA only)
ICD-Version	ICD-9	ICD-10
Timeframe	2001-2012	2019-2021
Use in this work	prototyping of pipelines	model development and tuning

As discussed in Chapter 2.1.2, developing a model with data from one hospital and applying it to another can be dangerous. For the software engineering work though it is advantageous to have real data that resembles the format of the ASIC data, while that cohort is still being collected. While with respect to modeling the risk of biases should be considered very thoroughly. Issues with different underlying practices and data structures exist between any two hospitals. These differences may get amplified when looking at hospitals further apart geographically. In addition to biases induced by, for example, the aforementioned changes in the disease landscape due to different climates, the medical procedures and practices change on a cultural and sociopolitical scale. With different policies on staffing, differing patient populations, levels of regionalization, and resources allocated, the extrapolations of any model with so significantly different train and test domains should be evaluated very carefully. For example, attitudes towards the end-of-life care can differ significantly. As such the responsibility for do-not-resuscitate

orders can be either made by the treating physician, or a patients' family or healthcare provider [132]. We in the end decided to exclusively use the MIMIC-III data for platform development and prototyping.

2.3 Data Imbalance

Considering a binary prediction task, if class A is significantly more common than class B, the dataset is considered imbalanced. The level of imbalance can range from slightly imbalanced dataset, for example the ASIC cohort with only 35% female patients, to a very strong imbalance, for example in fraud detection. An analysis from the German bank "Deutsche Bank" estimated 0.01% of transactions to be fraudulent [133]. The underrepresented class, female patients or frauds in these examples, is commonly referred to as the minority class, the other being the majority class.

This becomes a problem with machine learning, if the algorithms are overwhelmed by the much larger number of samples of the majority class and ignore the small class [134]. With canonical machine learning algorithms assuming a roughly similar distribution of classes, they will often end up biased towards the majority class. In many cases, the minority class will be the more important class from the researchers' perspective though, carrying important and relevant information. Considering machine learning in the medical domain, many machine learning tasks focussed on predictive modelling will be naturally imbalanced, as the events of interest occur rarely. For example in the research by Liu et al. [135], focused on predicting cerebral stroke. They used a public dataset with 43,400 recorded samples, out of which only 783 (1.8%) described stroke events.

When dealing with imbalanced data different metrics need to be considered when evaluating model performance. Accuracy, the overall percentage of correct predictions, for example, is misleading with imbalanced data. A predictor that always returns "this is not fraudulent" for every bank transfer would score an accuracy of 99.99% with the 0.01% estimated fraud rate presented previously. Such a predictor would naturally be useless in application. Different metrics, considering the minority class more carefully, are thus needed when working with imbalanced data.

		Prediction outcome	
		positive	**negative**
actual value	**positive**	True Positive	False Negative
	negative	False Positive	True Negative

Figure 2.2: Confusion Matrix Example

A confusion matrix (CM), as visualized in Figure 2.2, can be used to summarize the output of a prediction model. True Positive (TP) refers to samples correctly classified as true, for example fraudulent transactions, which actually are fraud. False Positive (FP) samples are incorrectly classified as fraud. They are normal transaction, which the model classifies wrongly. False Negatives (FN) refer equivalently to fraudulent transactions a model considers non-fraudulent. The True Negative (TN) transactions are those being correctly classified as non-fraud. For the CDSS which predicts a rapid loss of oxygenation, that will be presented in Chapter 4, TP are correctly raised alarms, with an event following. The TP describe when no alarm is raised, but also no alarm is needed. False negatives (FN) occur when an alarm should have been raised, but the system missed the event and false positives (FP) are false alarms, which occur when the model predicts an event to happen, but it does not. Based on this confusion matrix different metrics can be calculated, which incorporate the skewness in data distribution for imbalanced datasets better than accuracy. To reiterate, accuracy is calculated as the total ratio of true predictions, as shown in Equation 2.1.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

In the example of stroke detection, as inspired by Liu et al., a model never predicting stroke and labeling every event as the majority class (no stroke event), would still score an accuracy of 0.98 (refer to Equation 2.2). Due to the skewness of the data, the metric

no longer provides useful results.

$$accuracy = \frac{0 + (43,400 - 783)}{0 + (43,400 - 783) + 0 + 783} = 0.98 \tag{2.2}$$

With imbalanced data other metrics are often better suited, to reflect the interest in the minority class. Sensitivity and specificity are commonly used in the medical domain, with sensitivity (Equation 2.3) describing the ability to correctly predict true positives, and specificity (Equation 2.4) measuring the ability to avoid false positives. In the example of stroke prediction a highly sensitive model will be able to predict many stroke events correctly. If the specificity is high, the model would rarely wrongfully predict an event as a stroke.

Both metrics are very applicable to testing for the absence or presence of medical conditions. Considering, for example, a test for the presence of SARS-CoV-2, a high sensitivity indicates a high probability of detecting the virus, if it is present, while a high specificity indicates a high probability of a negative test result actually indicating the absence of the virus. In many cases there is a trade-off between sensitivity and specificity. An increase in sensitivity may decrease the specificity and vice versa; therefore it is crucial to always evaluate both measures.

$$sensitivity = recall = \text{true positive rate (TPR)} = \frac{TP}{TP + FN} \tag{2.3}$$

$$specificity = selectivity = \text{true negative rate (TNR)} = \frac{TN}{TN + FP} \tag{2.4}$$

Another important metric is the precision (Equation 2.5) of a model. While sensitivity relates the correctly predicted stroke events to all actually happening stroke events, precision focuses on the alarms raised by the model. Sensitivity is describing as the ability to correctly predict true positives. Precision in contrast is the fraction of relevant alarms in all alarms raised. To summarize, precision focuses on the accuracy of the alarms generated by the system, while sensitivity focuses on the ability of the system to detect events when they occur. Both are crucial when evaluating a model, especially for use in intensive care. Focussing only on sensitivity and specificity could lead to a system which, while being very good at identifying events and non-events, is not very accurate in its predictions. Consider

Figure 2.3: Confusion Matrix that would describe a model with high sensitivity and specificity, but low precision. Highlighting the relevance of paying attention to the right metrics in model evaluation.

		Predicted	
		0	*1*
Actual	*0*	99	28
	1	1	3

for example the confusion matrix in Figure 2.3. A model creating these predictions would score well for two of the three metrics. It would operate at $sensitivity = \frac{3}{3+1} = 0.75$ and $specificity = \frac{99}{99+28} = 0.79$, but it would generate many false alarms, with only 3 out of 31 alarms being relevant ($precision = \frac{3}{3+28} = 0.097$).

Finally, the F_1-Score (Equation 2.6), combines precision and recall (referred to as sensitivity earlier) into a single measurement, through the harmonic mean, providing a balance between the two. Including both, false positives and false negatives, this metric proves especially useful for imbalanced classes, allowing a holistic assessment of the models performance.

$$precision = \text{positive predictive values (PPV)} = \frac{TP}{TP + FP} \tag{2.5}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{2.6}$$

If a probabilistic model is used for binary prediction, a threshold needs to be set on the predicted probabilities to decide on the particular class the model predicts. If this threshold is crossed, the model will either classify the input as positive or negative. Adjusting this threshold changes the model's behavior and the associated metrics. A lower threshold may increase the sensitivity but could, in turn, increase the amount of false positives. The Precision is also influenced by this threshold, as it depends on the ratio of true positives to actual positives. A higher threshold may thus increase the precision but could lead to more false negatives. Selecting a higher threshold may also increase the specificity of the model, as the increase in true negatives improves the specificity but could, in turn, reduce the sensitivity of the model. With both variables used in the calculation

of the F_1-Score potentially being impacted by a change in the threshold, the score itself is also prone to change at different thresholds.

Various aspects influence the setting of this threshold, often depending on the domain and task at hand. In the medical domain, such influences could include clinical guidelines, cost-benefit ratios, or patient risk tolerances. Clinical guidelines established by medical professionals can define acceptable levels of sensitivity and specificity based on the diagnostic task or specific medical condition. The cost-benefit ratio refers to the potentially different consequences false positives and false negatives may have for different prediction tasks. Being able to prevent an upcoming life-threatening event by correctly predicting it, at the cost of a reduced model specificity due to an increase in false positives, may be desired in a clinical setting. This concept extends not only to the prediction task itself but to every single patient. Even considering the same model, e.g., for stroke prediction, a very different threshold might be chosen for patients, depending on their individual health status. For a more healthy patient, a model tuned to balance sensitivity and specificity might be a good choice, while a model tuned for high sensitivity might be a better fit for a critically ill patient.

A receiver operating characteristic (ROC) curve can be used to illustrate the performance of a model at different thresholds. The $sensitivity$ = true positive rate is plotted against $1-specificity$ = false positive rate, visualizing the performance achievable by the model. The area under the curve (AUC) for this ROC plot can be calculated (reference Equation 2.7), to achieve a single measurement to compare models across different thresholds. A random model would achieve, an ROC-AUC of 0.5, with a perfect model scoring 1.0. An example of ROC curves for three different classifiers are shown in Figure 2.4. The perfect classifier is shown at $(0, 1)$ achieving perfect $sensitivity$ and $specificity$. Three distinct classifiers are plotted, demonstrating how models can be compared. Every point of each curve refers to the performance of the respective model at one specific threshold for sensitivity and specificity.

$$ROCAUC = \int_0^1 \text{Sensitivity d(1-Specificity)} \tag{2.7}$$

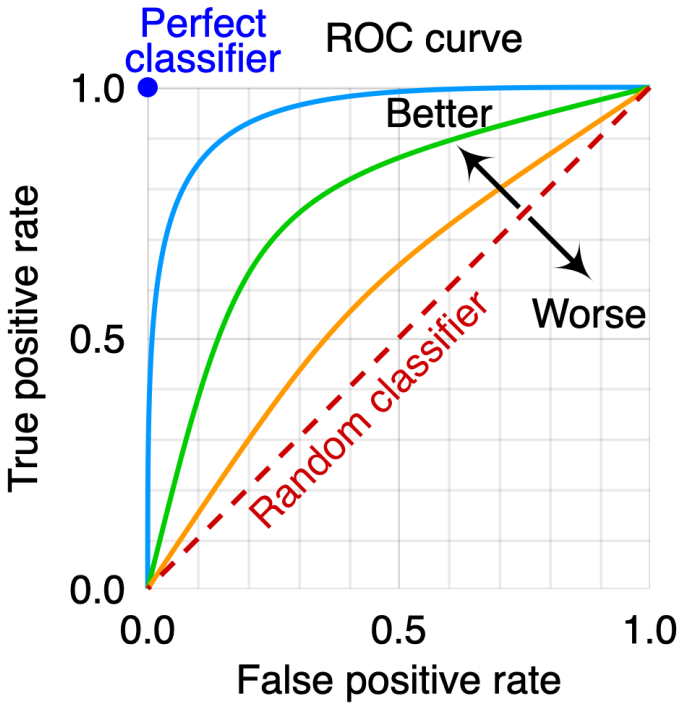


Figure 2.4: ROC curve example. A perfect classifier at (0.0,1.0) is indicated. The diagonal shows the performance of a random classifier. Three example classifiers are shown. Designed by MartinThoma and graciously donated to the public domain under CC-0 Public Domain license.

The ROC curve can mask poor performance in skewed data [136]. Research by Saito et al. [137] investigated the behavior of different metrics with imbalanced data. They recreated results from the MiRFinder study [138] and re-evaluated different classifiers comparing ROC curves and precision-recall curves (PRC) for five different models. The PRC is an alternative to the ROC curve, which still provides a wholistic overview of model performance at different thresholds. Figure 2.5 shows how the performance of most classifiers seems similar in the ROC space, but the precision-recall curves reveals bad performance of some classifiers. The corresponding AUC values are presented in Table 2.8. Most notably the RNAFold predictor achieves a high ROC-AUC at 0.964, comparable to the other high performing models, but the precision-recall curve reveals the performance to be significantly lower than the ROC-AUC indicates. The PRC proves also very useful in comparing the two best performing algorithms, MiRFinder and miPred, indicating a better performance from miPred, while both seem to perform identical in ROC-AUC space. Overall, the precision-recall curve is considered more informative when evaluating binary classifiers on imbalanced data.

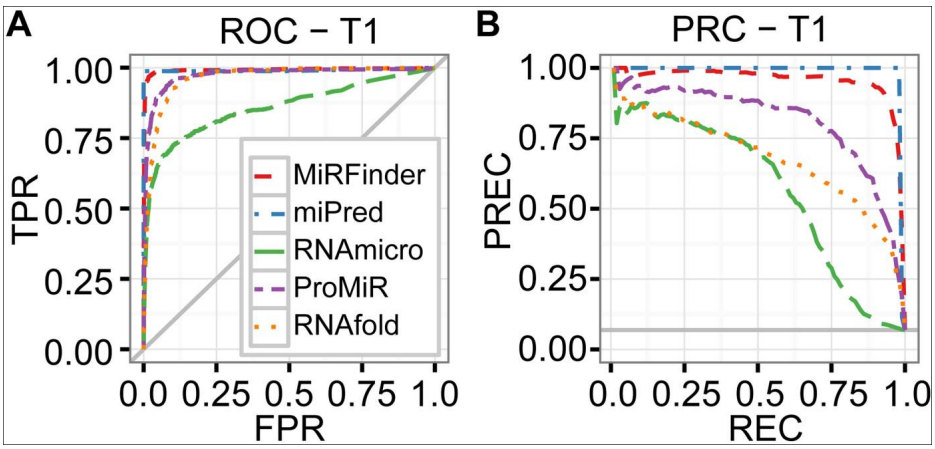


Figure 2.5: "A re-analysis of the MiRFinder study reveals that PRC is stronger than ROC on imbalanced data." ROC and PRC plots showing different predictors on the MiRFinder dataset. A gray solid line represents a baseline. The re-analysis is reproduced here from Saito et al. [137]

Table 2.8: AUC scores of ROC and PRC for T1. Reproduced from Saito et al. [137]

Model	ROC	PRC
MiRFinder	0.992	0.945
miPred	0.991	0.976
RNAmicro	0.858	0.559
ProMiR	0.974	0.801
RNAfold	0.964	0.670

As detailed earlier, for a CDSS alarm system, specificity describes how many alarms were raised out of all those that should have been raised, while precision describes how many of the raised alarms were relevant. The precision-recall curve, and the associated AUC, referred to as AUCPR in this work, thus more strongly incorporate the false alarms into the scoring. Especially in the noisy ICU environment, with alarm fatigue being a constant threat (refer to Chapter 1), this metric is highly relevant. With a focus on miss-classification of the minority class, this metric also works well with imbalanced data, as in contrast to, for example, accuracy, the minority class has a strong impact on the AUCPR.

To summarize, imbalanced data raises many challenges. Carefully selecting the right metrics for model evaluation is crucial, and the decision which metrics are right is

highly dependent on the specific task at hand. With the design of a CDSS alarm system for use in intensive care in mind, this work focuses on the F_1-Score and $AUCPR$ for model evaluation, while paying close attention to *sensitivity* and *specificity*.

2.4 Conclusion

This chapter summarized the datasets that were used during the research work described in this thesis. Two different datasets were used for different parts of the project. MIMIC-III, a freely accessible, large-scale critical care database, containing de-identified data from ICU patients from a single hospital in Boston, USA, was used for platform development and prototyping of model architectures. The ASIC project, a multicenter, observational study, spanning eight German University Hospitals, provided the data for model development and tuning. Due to data delivery delays, as well as potential biases incorporating different data from multiple hospitals, the data from only one hospital, the University Hospital Aachen, was used. The datasets contain patient vital signs, laboratory values, medications, and other relevant parameters collected during routine care in the ICU. Filtering through thresholds and a retrospective calculation of missing Horowitz index values were applied to the data to improve the data quality. Finally, this chapter discussed imbalanced data. Imbalanced datasets are common in machine learning, particularly in medical domains where events of interest occur rarely. To address this issue, different metrics need to be considered when evaluating model performance, such as precision-recall curves and the F_1-Score. These metrics will be used to evaluate the developed prediction model in Chapter 4.

Chapter 3

Diagnostic Expert Advisor

This chapter introduces the Diagnostic Expert Advisor (DEA), a model development platform in the context of medical timeseries data. The platform integrates HPC, data management, and visualization, focusing on ease of use and collaboration. It is designed to fill the need for research software at the intersection of ML research and clinical expertise. The platform presented in this chapter is used in Chapter 4, where a model for the prediction of rapid decline in oxygenation for intensive care patients is developed.

Chapter 3.1 introduces the need for such a platform and places it into the context of existing software solutions in this domain. Further, the design choices are explained, and the platform's architecture is detailed.

In Chapter 3.2, the individual features are highlighted and explained in more detail. Chapter 3.3 provides examples of the interface the user is facing when interacting with the DEA and showcases its use.

Finally, Chapter 3.4 highlights the impact of the DEA, and Chapter 3.5 summarizes this chapter.

3.1 A research platform for medical timeseries data

In Chapter 1, the need for good research software was highlighted. Different software frameworks with a focus on healthcare already exist. For example, the Medical Open Network for Artificial Intelligence (MONAI) provides a framework for medical data with a particular emphasis on imaging AI applications [139]. Another example is PyHealth [140], which provides standardized tooling for DL pipelines in healthcare applications.

Research on medical data utilizing AI-based technologies usually encompasses some form of data extraction, processing, and analysis at some point in their lifetime. These steps exhibit high similarity throughout projects, whereas application-specific adaptions of the workflow must be implemented, requiring standardization and flexibility. For example, data pre-processing might be necessary at load time or is executed once and saved to intermediate files. Rudimentary plots might suffice, or intricate interactive visualizations might be required. As the overall workflow of data extraction, processing, and analysis remains standardized but requires manual expert interaction, it can benefit from suitable software to enable better and faster research [141]. The DEA has been developed as a platform to start and homogenize research projects, providing standardization while encouraging customization to fit project and tooling requirements.

As a novel model development platform, the DEA's cornerstones are data management, visualization, and parallelization. The DEA offers sensible defaults, allowing for a much faster progression to model development by providing researchers with a platform with all steps already set up. Establishing a structure to adhere to offers various benefits. Researchers are encouraged to work more reproducible and organized while generating more easily shareable work. A unified platform further allows for easier reuse of existing code. For example, visualizations developed in one project can be quickly ported to another, as the fundamental data structure is similar. At the same time, through the DEA's flexibility, it is still possible to use the required or preferred tools, whether the project calls for a specific format for data storage, a preference for some visualization tool, or simply a dependency on an exact version of a machine learning library.

Another advantage of the DEA is the built-in support for parallelization. In

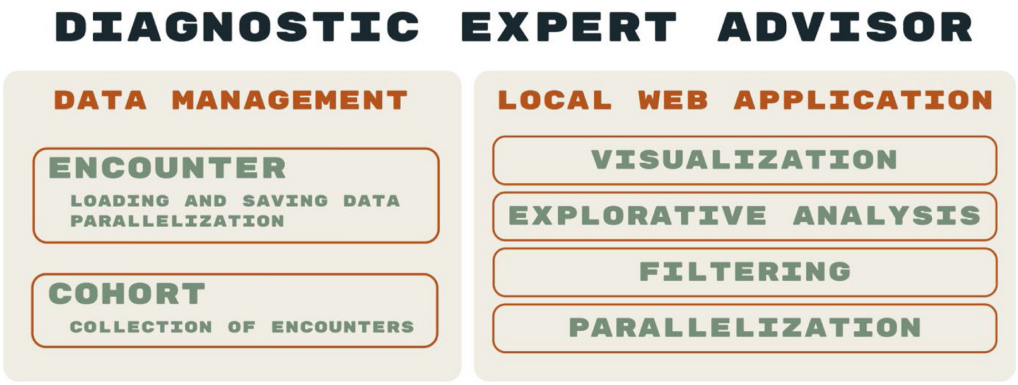


Figure 3.1: An overview of the architecture of the Diagnostic Expert Advisor. Data is managed through encounter and cohort classes. The local web application provides the interface for visualization, explorative analysis, and filtering, as well as the terminal to execute parallel calculations of HPC.

addition to local multiprocessing, the DEA integrates directly with HPC software. Researchers without an extensive IT background in data handling and analysis might be hesitant to use HPC resources available. The inclusion of parallelization-specific commands and structures can quickly result in convoluted code. The DEA separates those concerns by providing integrated and unobtrusive parallelization.

Both local multiprocessing and distributed HPC calculations can be achieved using the DEA. Researchers can write calculation methods as they would without parallelization, resulting in more readable, maintainable, and more easily shareable code. The DEA is tailored to the medical domain. While it can be extended to different areas, existing data structures assume a focus on heterogeneous timeseries data and a hierarchical organization of the data based on patient cohorts. It has been used to build a prediction model for the onset of acute respiratory distress syndrome (ARDS) in the SMITH project.

Various tools and platforms for tracking machine learning experiments have emerged. Examples include MLFlow [142], Neptune [143], Weights and Biases [144], and TensorBoard [145]. These programs often focus on tracking experiments and parameters for developing AI models. In contrast, the DEA is built to provide a platform for the fast integration and development of such models into the clinical context, filling a gap in the current research software landscape. A recent article described research software engineering as a pivotal and often undervalued research area and emphasized a need for infrastructure solutions allowing data to be made “interoperable, visualized and

leveraged by experts and non-experts alike" [146]. The DEA tackles these challenges at a small scope, enabling faster data comprehension and development of medical prediction models while still being accessible to medical experts and researchers without requiring an extensive background in software engineering.

The platform is built on Flask [147] as a local web app. This web app can be used as an interface for explorative analysis or to run code in parallel and on HPC hardware. Pandas DataFrames [148] are used for internal data storage. SLURM [149] is currently supported for interacting with the HPC, and the Joblib [150] library is used for intermediate data formats. Interactive visualizations are available through Bokeh [151]. PyGWalker [152] has been integrated as a Tableau [153]-like data exploration tool. Accommodating different choices for storage patterns or libraries has been a priority during the development of the Software. Therefore, all libraries, except for Flask, can be exchanged and adjusted to fit the needs of different researchers and projects.

3.2 Software Description

The DEA is written in Python [154]. In the standard implementation, data is saved to csv files, allowing easy interaction with other tools and languages. Such files can be opened by spreadsheet software like Microsoft Excel [155] or read in different environments like Matlab [156] or R [157]. The DEA is interfaced through a browser as a web application and the data management Python classes.

3.2.1 Software Architecture

As shown in Figure 3.1, the DEA consists of two components. A set of classes encapsulates the data to allow for parallelization, and a Flask server provides the web interface. The data management is based on wrappers around Pandas DataFrames. Thus, switching to different environments, such as Jupyter Lab [158], is possible without changing formats. The local web application is built on Flask, a well-established micro framework for building web applications. It can be customized in various ways. Some examples are explained in the next chapter.

3.2.2 Software Functionalities

The encounter and cohort classes allow for extensible visualization, data storage, and parallelization. An encounter describes an individual data point (in the context of our project: a patient's visit to the ICU). A cohort is a container for multiple of those encounters. They further allow for direct access to the underlying DataFrames, which can be utilized directly, e.g., for model development. Training of such models could then be run through the DEA directly on HPC infrastructure or entirely externally through other frameworks using the standardized Pandas DataFrames. Using the DEA requires transforming the data to this Encounter/Cohort format. Existing data needs to be imported into Pandas DataFrames. Pandas already provides many methods to read from CSV, JSON, Excel, or Apache arrow [159] files. A quick start guide [1] provides a sample implementation and explains these steps in detail. The functionalities outlined in this chapter are further described in the developer documentation [2]. All of them are designed to be customized and adjusted to the specific research task and data.

3.2.2.1 Visualization

Visualizations can be created in many ways. Due to the flexibility of providing a web server, almost all of them can be embedded into the DEA. Matplotlib [160] plots can be saved to static images and shown in plain HTML, while interactive plots, such as Bokeh visualizations, can utilize custom JavaScript. Plots can be defined on encounters and cohorts, providing different levels of detail. Cohort visualizations could include general population statistics, while encounter visualizations could focus on disease-relevant parameters and individual events in the ICU.

3.2.2.2 Explorative analysis

Through the visualizations, it is already possible to explore individual encounters with the DEA in detail. To further this capability and provide an effortless way to prototype new visualizations quickly, we integrated PyGWalker in the DEA for a tableau-like explorative data analysis environment. Through this component, researchers can interac-

[1] https://diagnostic-expert-advisor.readthedocs.io/en/latest/usage/quickstart.html
[2] https://diagnostic-expert-advisor.readthedocs.io/

tively create visualizations on the fly to test hypotheses or develop new ones. Plotting parameters without coding further encourages medical professionals and interdisciplinary researchers to investigate data.

3.2.2.3 Filtering

The DEA can load distinct cohorts, switching between ICU wards or hospitals. It is further possible to filter the encounters dynamically. Through filters, it is possible to explore different sub-cohorts of encounters and examine, e.g., especially endangered patients. Filters could also be used to evaluate model performance on different subsets of data. For example, in our project, we split the cohort into patients with varying lung failure levels to evaluate ARDS prediction models.

3.2.2.4 Parallelization

While many operations would benefit from parallelization, the increased effort during development often hinders its actual implementations. To encourage parallelization, the DEA provides a way to run calculations in parallel. These can be run either locally or on HPC hardware with limited setup.

3.3 Illustrative Examples

Upon starting the DEA, the user is presented a screen where a cohort can be selected. Afterward, an overview page is displayed. Figure 3.2 shows the various custom information for the cohort on this screen. It is possible to start calculations on the whole cohort through the "process" and "custom process" buttons. Both buttons trigger specific calculations on the HPC Cluster. The length-of-stay plot and all other information, like "Encounters in Cohort" and "Encounters Processed", can be fully customized. Badges can be added to provide a quick visual indication, e.g., that all data has been processed.

To inspect individual encounters, they can be searched through the navigation bar at the top, or an overview can be shown by clicking "Show Encounters". Figure 3.3 shows the corresponding UI. Custom information tags can be displayed per encounter to provide a quick overview. This allows, for example, encounters from extremely sick

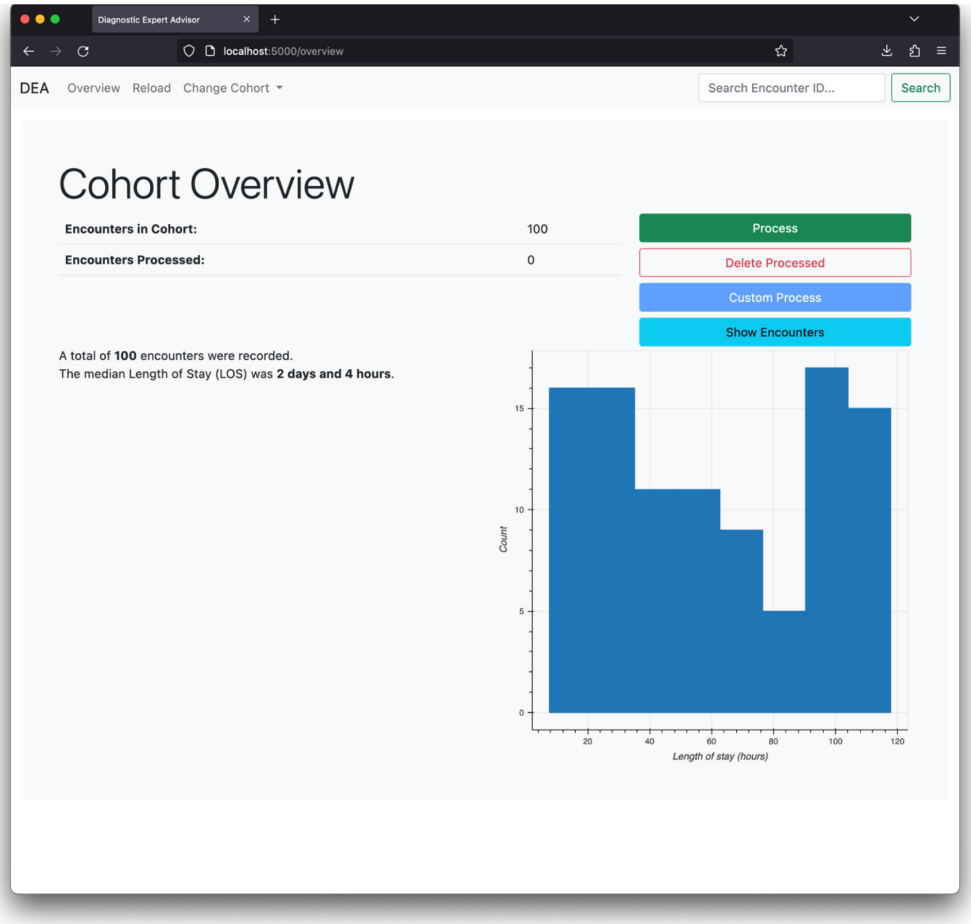


Figure 3.2: Cohort Overview, showing various customized elements. This figure illustrates, on the one hand, the customizability, as the LOS calculation, encounter stats, and the corresponding plot are all such customizations. On the other hand, the "Custom Process" button shows the capabilities to add custom calculations and processes to be run on HPC.

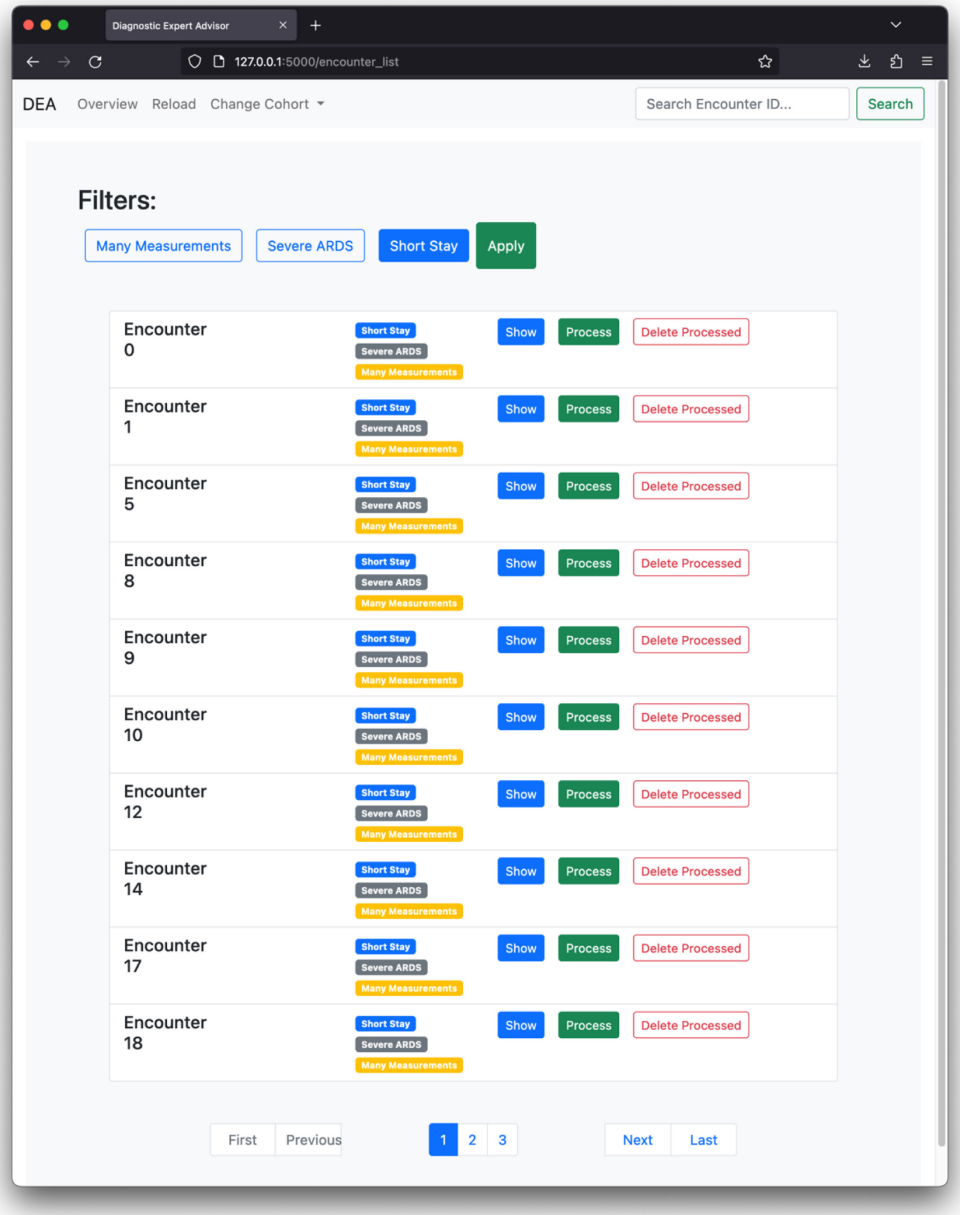


Figure 3.3: Encounter Overview, showing matches for the active filter "short stay". Filters, as well as tags, are fully customizable. This view presents the primary way to browse patient data in the DEA, allowing for fast visual identification of relevant patients through tagging and the execution of various pertinent commands during research and model development, such as re-running processing or removing intermediate files.

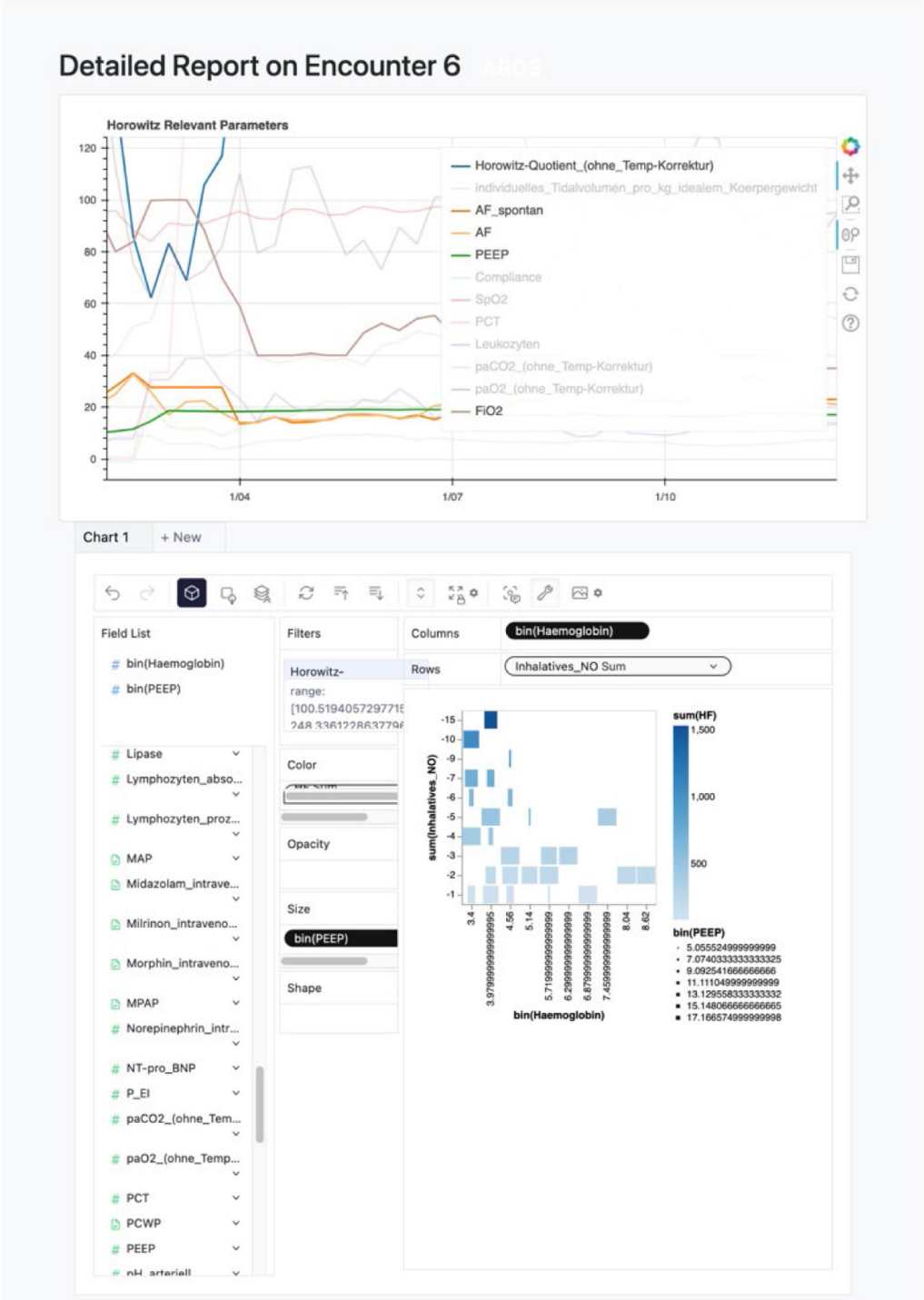


Figure 3.4: The default bokeh plot with ARDS-specific parameters and the PyGWalker interface is shown. This figure shows detailed information about a specific patient encounter in the ICU. It offers interactive plots to explore the data intuitively. The PyGWalker interface enables users to generate custom plots on the fly without touching code. This allows for a very rapid pace of data exploration and individual parameters to be explored and plotted against one another as needed.

patients or those with especially bad prediction results to be more visible. Filters for the selection of sub-cohorts are also available on this page. Calculations on individual encounters can be re-run, and all visualizations, including the PyGWalker Interface, are available when selecting a particular encounter, as seen in Figure 3.4.

3.4 Impact

The DEA provides a free and open-source research platform for developing AI models. As such, the DEA can improve the quality and speed of research work by promoting open collaboration, streamlining project setups, and enabling straightforward HPC utilization. Lowering the required expertise in software engineering allows medical professionals focusing on research to utilize such resources more quickly. Especially in medical research, an interdisciplinary joint effort between medical experts and researchers is crucial. In this unique environment, the DEA fills the need for standardized tooling, rapid prototyping, and reproducible and reusable research.

The design of the DEA has been steered by close cooperation with medical experts in the scope of the SMITH project [161] [162]. In the context of the Simulation and Data Lab Digital Patient, the DEA will be set up as a demonstrator in the National High-Performance Computing for Computational Engineering Sciences (NHR4CES) project [3], to showcase the ease of utilizing HPC resources to medical researchers. It will be facilitated and extended as a model development platform for the EDITH European Virtual Human Twin project [4].

3.5 Conclusion

In Chapter 1 the need for research software, specifically at the intersection of AI and healthcare, was highlighted. The DEA sets out to solve this problem by providing an open platform for developing machine-learning models on heterogeneous medical timeseries data. It enables researchers to use a streamlined workflow and encourages a replicable and organized setup. Making parallelization and interactive visualizations

[3] https://www.nhr4ces.de/simulation-and-data-labs/sdl-digital-patient/
[4] https://www.edith-csa.eu

more approachable supports the more widespread adoption of these features. The innate data structure and the ability to quickly investigate individual encounters foster patient-centric research. The platform provides a foundation for many medical research areas while staying generic enough to be adaptable to other domains. It is designed to be flexible and adaptable. Whether data is extracted directly from a hospital database or cross-referenced from another project, both approaches can be accommodated with few changes to the DEA. Implementing the DEA as a platform encourages more organized, readable, and shareable research.

Chapter 4

Predicting rapid loss of oxygenation intensive care patients

This chapter describes the development of a predictor for rapid decline in oxygenation in the context of ARDS in the ICU. The platform developed in Chapter 3 provides the framework for developing this ML model on the ASIC data detailed in Chapter 2. In this chapter the data processing is detailed, a surrogate marker for ARDS is defined, and the model setup is described. Then the results are presented and compared to the literature, and the model's ability to generalize is evaluated by applying a model not trained on any SARS-CoV-2 patients to predict ARDS in patients infected with the virus. Finally, the results and limitations are discussed.

Chapter 4.1 briefly summarizes the motivation for the prediction of rapid decline in oxygenation. It further places this work into context and elaborates on the current state of the art with respect to the prediction of ARDS onset.

Chapter 4.2 describes any model-specific processing to the data not already detailed in Chapter 2. Most notably, the transformation of the original timeseries data into

a format fit for the efficient digestion by XGBoost, and the extraction of two distinct cohorts, stratified based on the presence of SARS-CoV-2.

In Chapter 4.3, the definition of ARDS is reviewed, and its applicability within the ASIC project is explored. A clinically relevant surrogate is defined, inspired by both clinical practice and the Berlin definition diagnosis criteria.

Chapter 4.4 describes the prediction model used for this work, provides some relevant background knowledge, and outlines the training and validation process.

Finally, Chapter 4.5 details the results with dedicated sections on both the generalizability of the model and the impact of different sampling strategies for the imbalanced learning problem, as introduced in Chapter 2.3. The results are then discussed and placed into context in Chapter 4.6.

4.1 Introduction

Chapter 1 introduced ARDS, a dangerous syndrome where the lung can no longer oxygenate the bloodstream sufficiently. In critically ill patients, ARDS is a common cause of respiratory failure [163]. With roughly 10% of patients admitted to ICU either being affected by ARDS at admittance or developing the syndrome throughout their hospitalization, ARDS is a major burden on global healthcare. No pharmacotherapies for ARDS have yet been identified. A recent study evaluating aspirin as a preventive drug for at-risk patients did not find any significant effect of the treatment [164]. Current treatment is thus focused on lung-protective ventilation. Research suggests that, in mechanically ventilated patients, a conservative fluid management strategy improved lung function and shortened the MV and ICU duration for patients [36]. Early identification of patients developing ARDS is crucial to implementing lung-protective ventilation and deploying a fluid-conservative approach. Different clinical scores for predicting ARDS patients have been developed, such as the Lung Injury Prediction Score (LIPS) [165]. The overall performance of LIPS is modest, though, when applied in settings outside the original hospital environment [166] [167]. Further, it can be outperformed by ML-based models [10]. In contrast to clinical scores directly derived from expert knowledge, ML methods

can be used on the data available for ICU patients to predict if a patient is at risk of developing or progressing into a more severe form of ARDS.

Predicting ARDS is an active area of research encompassing different approaches and strategies. Researchers focus on various prediction targets, such as mortality [168] [169], mechanical ventilation duration [170], ARDS severity [171], or ARDS presence from radiology reports [172]. The motivation for this work is the prediction of the onset of ARDS. With the data collected from the ASIC study, evaluating the full Berlin definition for ARDS onset is impossible, as elaborated on in Chapter 4.3. A comparison to predictors that target ARDS onset still offers context concerning state of the art, even if the prediction target differs. Taoum et al. [8] developed a prediction model based on novelty detection and data fusion. They achieved a sensitivity of 65% and a specificity of 100% on average 39h prior to onset, using continuous physiological signals of heart rate, respiratory rate, peripheral arterial oxygen saturation, and mean airway blood pressure. Their work, however, was only tested on a small dataset with samples measured every minute. This may impact the generalizability of the approach and might limit its applicability in other hospitals where measurements aren't available at such high resolutions. A different approach, by Zaglam et al. [173], utilized Linear Discriminant Analysis (LDA) and chest radiographs, showing a sensitivity of 90.6% at 86.5% specificity for the detection of ARDS in the radiographs, where 53 out of 90 images in the test set presented ARDS. Le et al. [9] deployed an XGBoost gradient boosted tree model, the same model used in this work, on MIMIC-III data, to predict ARDS onset based on the Berlin definition. They further include a Word2Vec [174] representation of radiology text reports, which are not available in the ASIC dataset. They report results for prediction windows of up to 48 hours and achieve results that outperform other studies. The cohort used was not restricted to mechanically ventilated patients, which is often a requirement in other studies [8] [175]. They achieve a ROC AUC of 0.90 at onset time, and 0.79 for predicting 48h prior to onset, with a respective sensitivity of 0.67 and a specificity of 0.85 with the 48h prediction window. Singhal et al. [10] present an ML algorithm called "eARDS" in their research, predicting ARDS in COVID-19 ICU patients. The model predicts the onset of ARDS based on the Berlin definition up to 36 hours before the event in a multi-cohort study encompassing data from ICUs at Emory Healthcare, Atlanta, GA, the University of

Tennessee Health Science Center, Memphis, TN, and the Cerner® Health Facts Deidentified Database. Overall, more than 35 thousand ARDS patients were used in the model. They achieve a ROC AUC of 0.89 at a sensitivity of 0.77 and a specificity of 0.85 using a XGBoost model as well.

To our knowledge the studies by Le et al. and Singhal et al. both defined the onset time t_{onset} as the first timepoint where the PaO_2/FiO_2 ratio fell below or equal to 300 mmHg while a PEEP of above or equal to 5 cmH_2O was present. Singhal et al. further stratifies the severity of ARDS based on the lowest PaO_2/FiO_2 ratio observed throughout the patient stay. An overview of the different models is presented in Table 4.1.

Table 4.1: Comparison of predictors developed for ARDS onset in literature. The annotation Berlin* denotes that the authors focus on a Horowitz Index $\leq$ 300 while PEEP $\geq$ 5. Le et al. further include the radiology report. None use expert annotation to comply with the full Berlin definition.

Authors	Dataset	Model Type	Horizon	Target	Best ROC-AUC	Sensitivity	Specificity	Notes
Zaglam et al.	TARD	LDA	—	ARDS	—	0.91	0.87	Detect ARDS in Radiograph
Taoum et al.	MIMIC-II	Novelty	Up to 39h	Berlin*	0.79	0.65	1.00	Small Cohort, HighRes data
Le et al.	MIMIC-III	XGBoost	Up to 48h	Berlin*	0.90	0.67	0.85	Includes Radiology Reports
Singhal et al.	Multicenter	XGBoost	Up to 36h	Berlin*	0.89	0.77	0.85	Focus on COVID patients

4.2 Data Processing

This section refers to the ASIC dataset, described in Chapter 2.1, and elaborates on the aspects relevant to model development. Overall, 81 parameters from the ASIC study have been used in this research. They can be separated into dynamic parameters and static parameters. Static parameters include, for example, *age* and *sex*, which are usually recorded only once during a patient's stay at the ICU. Dynamic parameters refer to all parameters measured over time, such as *heart rate* or *lab values*. An overview of the parameters used can be found in the Appendix. The total amount of visits to the hospital data collected during the study is 3,676. Due to data privacy concerns, data has been

exported in a way, such that is impossible to discriminate between multiple visits of the same patient to the ICU ward. Thus, this work focuses on individual visits to the ICU, referred to here forth as encounters, and will not differentiate by patients. For the 3,676 visits to the ICU, the mean length of stay was 20.5 days, with a standard deviation of 21.8 days. The shortest recorded stay was around four and a half hours, and the longest was almost 187 days. The 25% quartile has a length of stay of 6 days, the 50% quartile has about 12 days, and the 75% quartile ends up at 25.5 days.

Missing data was filled using last-one carry forward imputation, replacing missing values with the last valid data point. Most of the data worked with remains "valid" until a new measurement is recorded. For example, in ventilator settings, many hospital systems only register the settings when set, but they remain the same until a new parameter is recorded. Forward filling is commonly used in timeseries prediction tasks, as it ensures the data integrity is kept, not allowing temporal distortions to manifest. Other imputation techniques, such as interpolation, could leak information that is available only in the future into the past. Static parameters are, due to the study design, clustered for k-anonymity. Where applicable, they were included in the training data and encoded numerically. ICD codes are utilized to define the different cohorts retrospectively. They are not included in the data available to the model, as they are usually unavailable throughout the hospital stay. Dynamic parameters are exported at a maximum resolution of 15 minutes, with many parameters, such as lab values, being recorded less regularly.

Listing 4.1 ICD-10 Codes that are considered for the stratification of cohorts in the ASIC dataset.

```
U07.1: COVID-19, virus identified
U07.2: COVID-19, virus not identified
U07.3: Previous COVID-19 infection affecting the patients health
U07.4: Post-COVID-19
U07.5: Multisystem inflammatory syndrom, related to COVID-19
```

Three cohorts were created from the dataset. One cohort contained all data and was used to establish a generic model. Two additional cohorts were used to test the generalizability of the model with respect to SARS-CoV-2. The ICD Code *U07* was used to stratify the patients. In the data available for this project this ICD Code was most reliably used to annotate a SARS-CoV-2 [176] infection. The codes considered are detailed in Listing 4.1.

Table 4.2: Overview of the cohorts created for model training and testing.

Name	Description	Encounters	Total Timepoints	Event Rate	ARDS Prevalence
FULL	Contains all encounters available	3676	7,237,099	11.2%	21.7%
NON-COVID	Encounters without the presence of SARS-CoV-2 as Encoded by ICD U07	3380	6,343,647	11.9%	17.2%
COVID	Encounters with U07.01 or U07.02 ICD-Codes present, denoting an infection with SARS-CoV-2	296	893,452	6.8%	84.1%

U07.1 denotes a case with a virus load verified by a laboratory test, and *U07.2* indicates an assumed infection without lab verification yet. *U07.3* and *U07.4* were only used four times in total and did not necessarily denote acute SARS-CoV-2. The cohort of infected patients, denoted as the *COVID* cohort, contains 296 individual ICU visits. The remaining 3,380 visits are part of the *NON-COVID* cohort. Table 4.2 provides an overview of the data availability, event rates, and prevalence of ARDS in the two cohorts defined. In Table 4.3, the distribution of ARDS and the percentage of ARDS patients in the whole population, stratified by severity, is provided. The *COVID* cohort shows a significant increase, especially in moderate and severe ARDS. Table 4.4 shows general population statistics for the three cohorts. Patients in the *COVID* cohort are more likely to be younger and male, they usually stay longer, as visualized in Figure 4.1, and more likely to die in hospital.

Table 4.3: Overview of respective ARDS severities present in all cohorts. Providing total counts of encounters with the ARDS severity as denoted by the ICD code and the percentage with respect to the total cohort size.

Name	Mild ARDS	Moderate ARDS	Severe ARDS
FULL	52 (1%)	342 (9%)	437 (12%)
NON-COVID	46 (1%)	263 (8%)	273 (8%)
COVID	6 (2%)	79 (27%)	164 (55%)

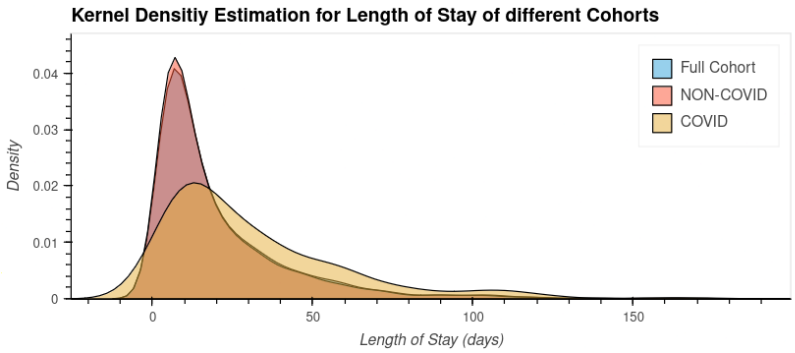


Figure 4.1: Kernel Density Estimation (KDE) of the length of stay for the three different cohorts.

Table 4.4: Overview of population statistics for the three defined cohorts.

Name	< 70 years	70-79 years	> 79 years	Male	Length of stay (mean±std)	Mortality
FULL	62%	25%	13%	65%	20.5 ± 21.8 days	33%
NON-COVID	60%	26%	14%	65%	19.5 ± 20.8 days	32%
COVID	75%	20%	5%	69%	31.3 ± 28.5 days	44%

Training and test data is derived by transforming the data from the original timeseries format to a format better suited for the deployed XGBoost model. Figure 4.2 illustrates the concept of this transformation. The data from a historical window is described through some statistics in a rolling window fashion. The final model used min, max, median, mean, std, as well as the difference between the last and first values in the window. This description reduces the dimensionality of the input data from 24 hours per window, each hour consisting of four 15-minute intervals, making for a total of 96 features per parameter, to merely six. Static parameters are appended to the resulting vector and provided to the model at every time step. While this approach is not memory efficient, it is both practical and easy to implement [177].

4.3 A surrogate marker for ARDS

The Berlin definition is commonly used to diagnose ARDS nowadays. It improved the previously established AECC definition and currently represent the most widely accepted definition of ARDS. The diagnostic criteria are listed in Chapter 1. Most notably,

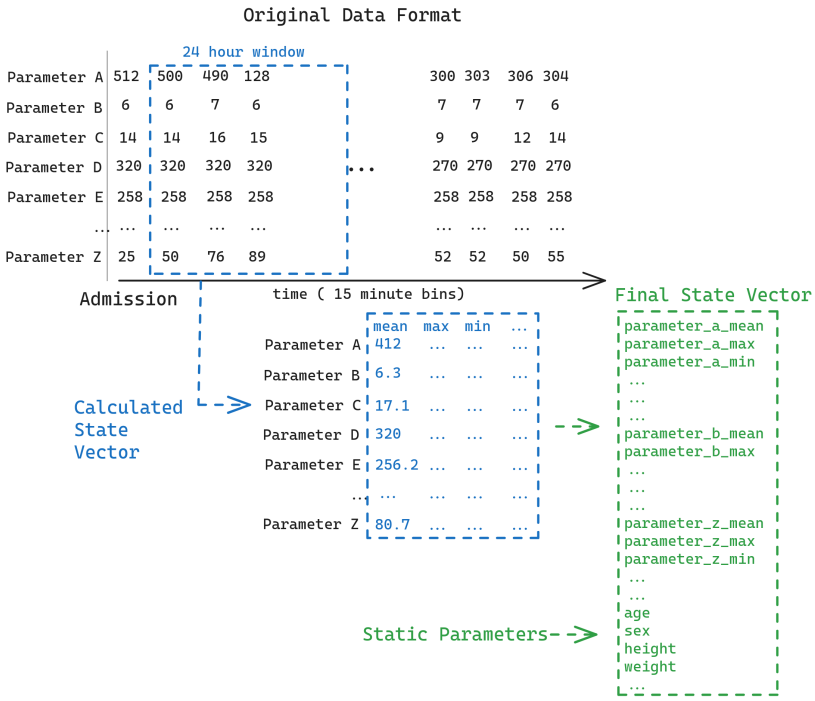


Figure 4.2: Visualizes the data structure for model training. Past data for all parameters is described through different statistics. The resulting matrix is transformed to a vector and the static parameters are appended. This vector is then used with the XGBoost predictor.

this definition requires medical experts to exclude heart failure or volume overload as a reason for a low Horowitz index value, and an examination of a chest radiograph or computed tomography (CT) image. As the dataset does not include any chest imaging, and alternative explanations for respiratory failure are hard to derive from data without explicit labelling from medical experts, the model we develop focuses on a decreased Horowitz index to predict the onset of ARDS. Approaches focusing on the Horowitz index value are commonly used in this case, as detailed in Chapter 4.1.

While approaches found in literature often closely adhere to the Berlin definition, this work focuses on a rapid decline in Horowitz index. We define a binary marker based on the median Horowitz index over two 24-hour intervals, to reduce the impact of erroneous measurements and better depict the overall trajectory of the lung health. Considering the current patient state, the first interval extends 24 hours in the past, with the second interval extending 24 hours in the future after some prediction horizon. The final model uses a 48-hour horizon, predicting how the Horowitz index will change at prediction time t from the previous day $[t - 24h; t]$ to a window ending three days in the future

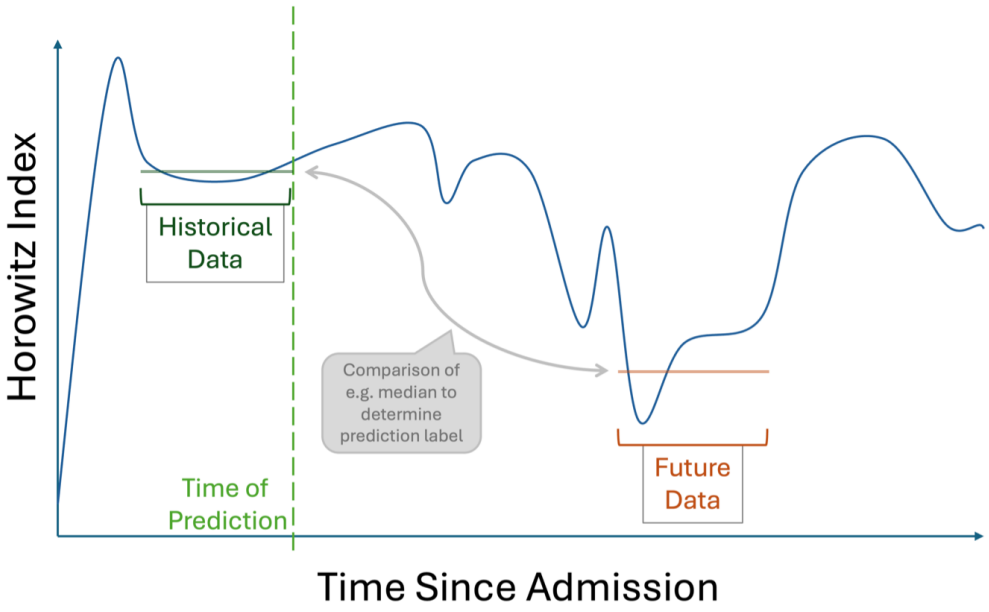


Figure 4.3: Visualization of the Surrogate Marker for ARDS. The data from the historical data window is compared to that of the future data window to determine whether an alarm should be raised or not.

$[t + 48h; t + 72h]$. Figure 4.3 below visualizes the described concept.

The binary marker was considered active when one of the following criteria is met:

- `(a) Horowitz index was over 450 and dropped more than 100`
- `(b) Horowitz index was over 350 and dropped more than 60`
- `(c) Horowitz index was over 250 and is now under 200`
- `(d) Horowitz index was over 150 and is now under 100`

The above thresholds have been defined in collaboration with medical experts to capture the dynamics of a decreasing pulmonary function. Criteria `(a)` and `(b)` define a change as significant if it is larger than 100 and 60 respectively, based on whether the Horowitz index in the historical window was larger than 450 or 350. A larger change in Horowitz index is required, if the oxygenation was better initially. Such events would often indicate rapid onsets of ARDS, with a quickly worsening patient state. The definition of criteria `(c)` and `(d)` are leaning on the Berlin definition for their thresholds, but instead of focusing on fixed values, they again prioritize a rapid decrease. This approach in general

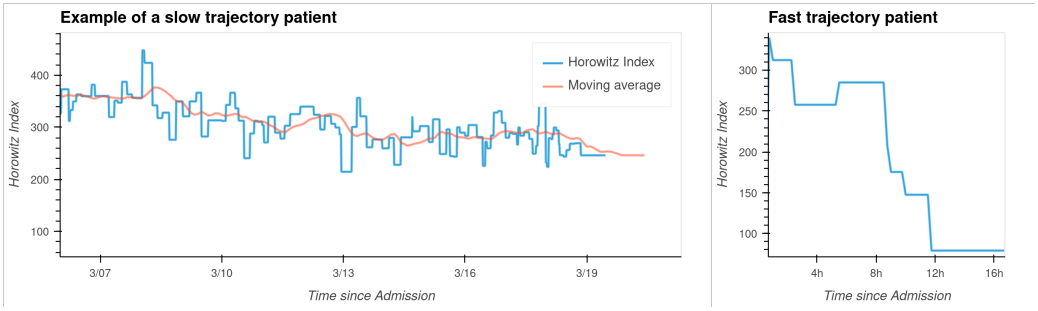


Figure 4.4: Comparison of different trajectories in Horowitz index decreases. The left plot shows a patient with a slow trajectory, the Horowitz index decreasing from around 350 to around 250 over 12 days. In the right plot the Horowitz index drops to less than 100 from previously over 300 in roughly 12 hours.

has two advantages over fixed thresholds often used in literature. On the one hand, a patient hovering around a threshold, repeatedly measuring above then below it or vice versa, would not cause multiple consecutive alarms to be sounded by our model. On the other hand, patients with a slow and steady declining Horowitz index are omitted by the model with this definition, allowing the development of a model focussing on rapidly deteriorating patients.

Concerning ARDS, we roughly categorize two different trajectories of syndrome progression for a CDSS. If the patient's condition changes slowly and consistently, a CDSS detecting trends and monitoring thresholds should be capable of properly capturing the dynamic and then providing proper alarms. Patients with such temperate and consistent trends will need less immediate care due to their slow trajectory. Predicting those might be possible with a knowledge-based CDSS that regularly checks the Berlin criteria. A rudimentary version of such a system was developed as part of the ASIC study through an app, that encourages conscious checks when the thresholds defined by the Berlin definition are met. If, on the other hand, the change in Horowitz is neither slow, nor consistent, but is instead volatile and sudden, an ML-based CDSS is more suited. Utilizing the vast amount of data generated in intensive care environments, such a system could pick up underlying patterns and predict rapid declines in a patient's state and thus, for example, a sudden onset or worsening of ARDS. Figure 4.4 illustrates this on two examples from the ASIC cohort. In the left plot, a patient is slowly getting worse over twelve days, with the average Horowitz index decreasing from around 350 to 250 in the end. An Alarm when the patient crossed the threshold of 300 would suffice to alert the medical experts

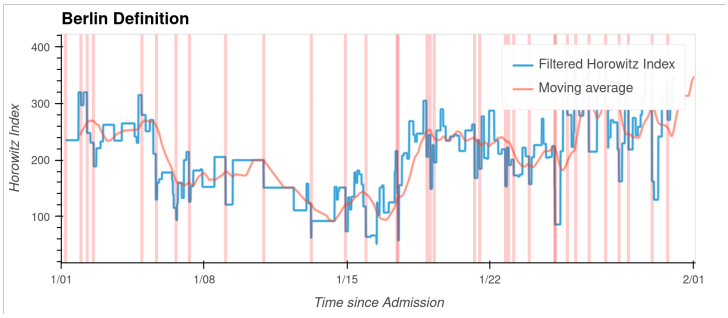


Figure 4.5: Event that would cause alarm for the marker derived from the Berlin Definition. Horowitz index drops below 300/200/100 each would cause such an alarm, indicated by a vertical line.

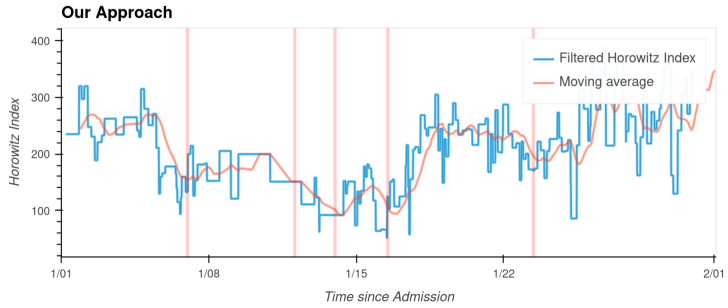


Figure 4.6: Event that would cause alarm for our marker. Alarms are indicated by a vertical line.

to the overall decreasing patient health. The right pot on the other hand shows a patient with a Horowitz index decreasing from over 300 to less than 100 in roughly 12 hours.

Figure 4.5 shows time points at which the Horowitz index drops below the fixed thresholds defined in the Berlin definition. This approach, often in combination with the $PEEP \geq 5$ criteria we omit here for the sake of illustration, is commonly used in literature. Every vertical bar denotes an event a model would usually predict. Our approach in contrast is focused on the difference in the median Horowitz index in two prediction windows, as visualized in Figure 4.3. Figure 4.6 shows the same patient with alarms generated by our approach. For this specific patient five relevant area were identified, indicated by the vertical bars in the lower part of the plot. The moving average over a 24-hour window calculated every two hours is overlaid onto the filtered Horowitz Index to better visualize changes in the underlying dynamic. The approach derived from the Berlin definition identifies 34 distinct events overall, based on thresholds crossed. While undoubtably the Horowitz index thresholds are crucial for the correct diagnosis and the assessment of ARDS severity, we hypothesize our marker to be clinically more relevant in

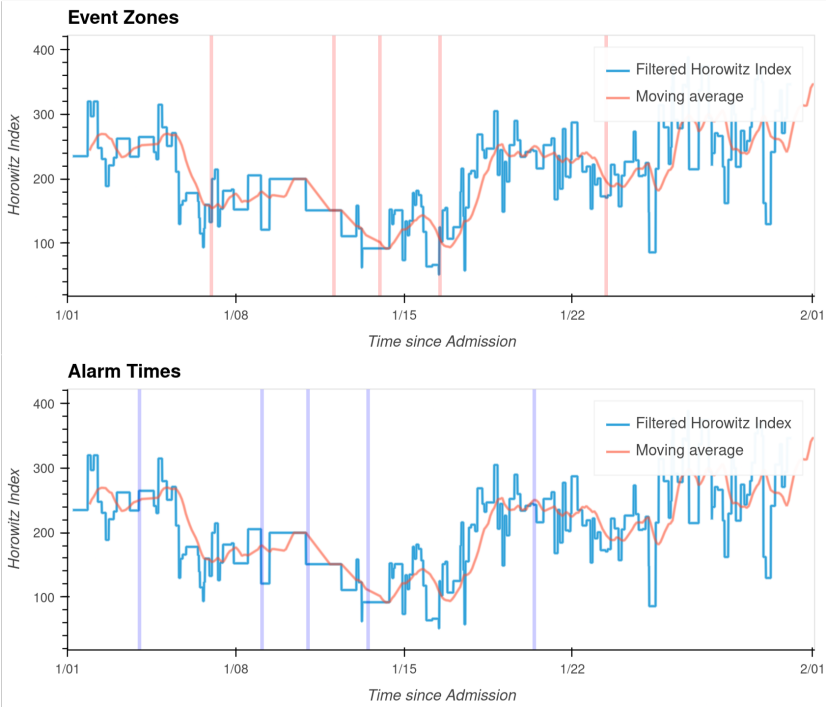


Figure 4.7: Visualization of relevant zones identified and the corresponding alarms three days prior.

the context of CDSS alarm systems in the ICU.

Figure 4.5 and Figure 4.6, both visualize the events of interest. For the prediction model we developed a horizon of up to three days is feasible. Figure 4.7 illustrates for our markers the times at which alarms would be generated for this example.

Table 4.5 describes how long it takes until the Horowitz index falls below the threshold of 300/200/100, as used to describe the severity of ARDS in the Berlin definition. Unsurprisingly, most patients in intensive care quickly drop to a Horowitz index ≤ 300, with an average of 23.6 hours until this threshold is crossed. While mild ARDS, as defined by the highest threshold of 300, is often present early on, the more severe stages of ARDS take longer to develop.

Figure 4.8 visually compares the time it takes until a threshold is crossed and shows, that progressively more severe Horowitz index states develop later in the ICU stay for most patients. On average, if a Horowitz index below 100 is reached throughout the stay in the ICU, it happens after 94.1 hours.

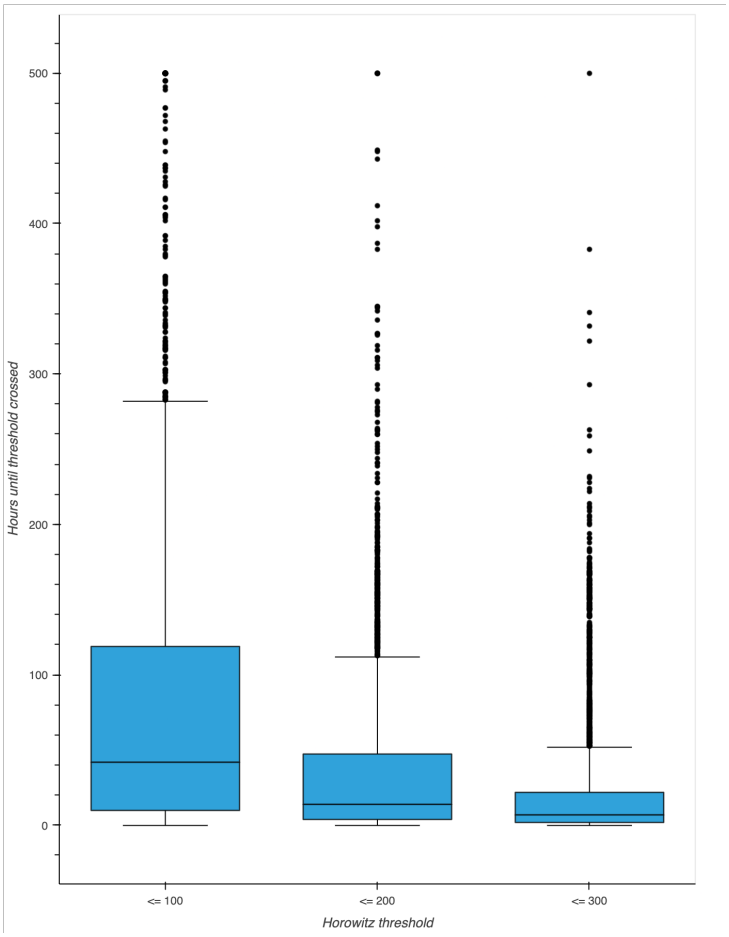


Figure 4.8: Visualization of the time it takes until a specific Horowitz index threshold is crossed. The plot is clipped to 500 hours max.

Table 4.5: Overview of average times until Horowitz index first drops below the thresholds defined in the Berlin definition.

Horowitz index below	Count	Mean	Std	25% Quartile	50% Quartile	75% Quartile
100	2197	94.1 h	151.6 h	11 h	43 h	122 h
200	3469	37.7 h	60.2 h	4 h	14 h	47 h
300	3639	23.6 h	42.9 h	2 h	7 h	22 h

4.4 Prediction model and pipeline

This work uses gradient boosting of regression trees, originating from the research of Friedman et al. [178], as implemented in the XGBoost [179] framework. XGBoost employs decision tree ensembles containing classification and regression trees (CART) [180]. In those trees, leaves are assigned to numerical values instead of classification results, allowing for more detailed interpretability of the ensemble and a unified optimization approach.

Considering the generic supervised learning task where a prediction y_i is made from input x_i the ensemble of trees can be written as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

where K denotes the number of trees in the ensemble. $\mathcal{F}$ is the function space of all possible CARTs and f_k is a single function from that space. Trees in this ensemble are trained in an additive manner, adding additional trees at every training step. The core idea is that every new tree improves on the errors of the combined ensemble of existing trees. The implementation through XGBoost is highly effective and scalable, includes both Lasso and Ridge regularizations to combat overfitting, and has built-in capabilities to handle missing data. The framework is well documented, and the ensemble models are fast to train, memory efficient, and perform well. For example, XGBoost models regularly achieve high rankings in prediction competitions on the data science platform Kaggle [1] [181] [182] [183], across different disciplines. They are utilized in various medical

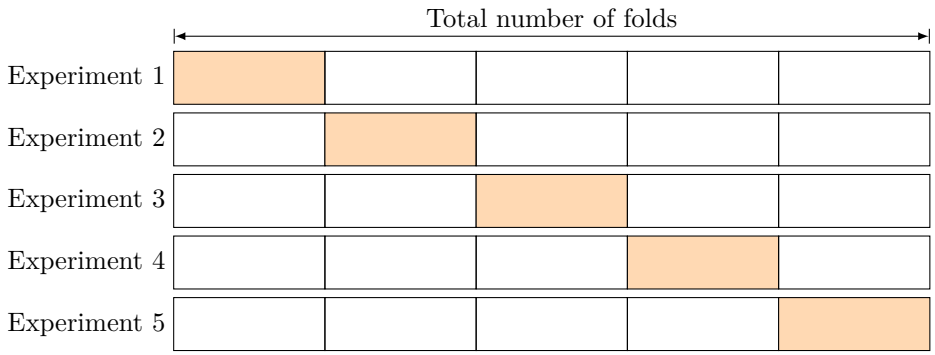
[1] https://www.kaggle.com

prediction tasks as well, often outperforming alternative approaches in both ease of use and predictive performance [184] [185] [186].

When developing machine learning models that learn from data, it is of crucial importance to carefully separate which data is provided to the model at which state. Data is usually divided into multiple data sets for this reason. Most commonly three different sets are used during model creation: training data, validation data and test data. The model is developed and validated with the training and validation data. Hyperparameters of the model can be tuned, manually or algorithmically, to ompitimize the performance of the model on the validation data. As the model is in turn being optimized for the prediction of the validation data, the generalizability of the model to data never seen before is finally tested using the test data. With the test data not informing any part of the model, either directly (training data) or indirectly through the hyperparameters chosen (validation data), this test data is used to derive the final model performance. When creating these train/test/validation data sets it is important that they represent the overall dataspace well. One way to ensure the individual sets do not vary videly is by stratifying for the prediction target. An additional consideration that needs to be taken into account arises from the temporal data in this work. In contrast to independent data, which can often be freely shuffled, split, and merged, temporal data can leak information between test, training, and validation, even if the data is split into different sets. To ensure data from individual encounters is not dispersed over different data sets, we limit the splitting in such a way, that a single encounter can only be part of either the training, testing, or validation set. To summarize, the available data is split into training, testing, and validation sets, while ensuring that the over all ratio of events is similar in all sets, and further no encounter can be split into different data sets, such that temporary related data is kept in the same set.

Many machine learning algorithms are non-deterministic. They can exhibit different behaviors on different runs. Neural networks, for example, can be intialized with random weights at every run. When evaluating a machine learning model cross-validation is often used to derive more acurate estimates of the predictive performance of the model [187]. This is achieved by repeatedly training and testing the model on different parts of

Figure 4.9: Cross Validation Visualization



the data, while avoiding problems like overfitting and selection bias [188], similar to the data splitting detailed above.

For this work k-Fold cross-validation (CV) was used to approximate the model performance. In k-fold cross-validation the data is divided into equally sized sets. The k denotes the number of sets, usually referred to as "folds". These folds are then used to train and test the model, with some folds used for training, testing, and validation each. Considering the 5-fold cross-validation in Figure 4.9, a model would be trained on all but the highlighted folds and then be evaluated on the left out fold. Overall 5 different models would be trained and evaluated thus, and the performance of the models could be averaged over the different runs. For the data this work is concerned with, it is once again crucial to ensure that individual encounters are not spread out over different splits. It is furthermore still desireable that each split has both, a roughly similar amount of total data, as well as a roughly similar event rate, to all other splits.

This work uses the pipeline described in Figure 4.10. The data is prepared as described in Chapter 2 and Chapter 4.2. Then it is split into training and test data, with 10% hold-out test data and 90% of the data used for training. Hyperparameters are tuned on the training data using CV. The tuning was run through Optuna [189] and utilized bayesian hyperparameter optimization. The bayesian approach allows for a efficient exploration of the hyperparameter space. Instead of evaluating every combination of hyperparameters informed choices are made, often leading to a faster convergence to optimal, or near-optimal hyperparameters [190]. The list of hyperparameters explored is available in the Appendix. One thousand trials were run, and the performance was

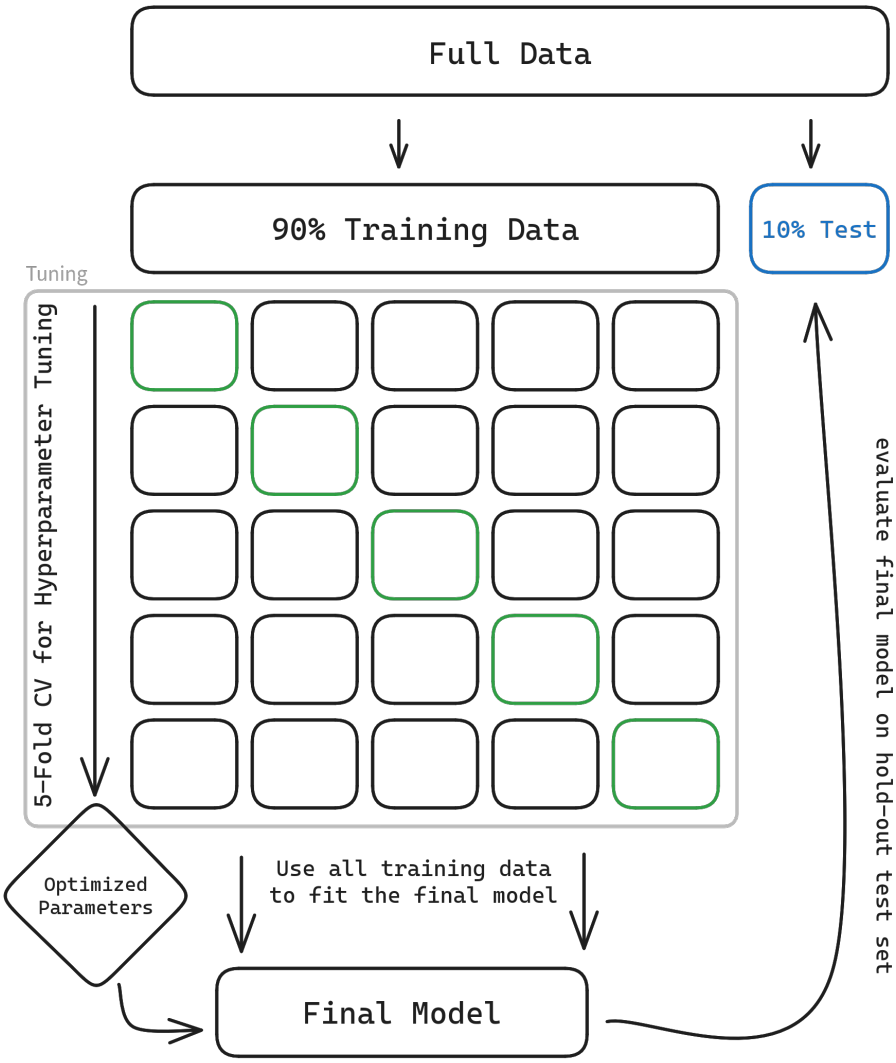


Figure 4.10: Pipeline used for model development in this work.

evaluated through 5-fold cross-validation, stratified for events and ensuring encounters are only part of one fold. Early stopping was used to cut off unpromising trials if the median result was worse than the median result of previous trials at the same step. The model was optimized for the Area under the Precision–Recall curve. Results were logged in MLFlow [142], and the Diagnostic Expert Advisor [12] was used for data handling and processing. All code was developed in Python [154], using SciKit-Learn [191] for many of the ML related tasks.

After optimal parameters had been found, a model with those parameters was trained at a reduced learning rate and extended training rounds until convergence on all

training data. This model was then used to evaluate the final performance on the holdout test set. For the evaluation of the generalizability to SARS-CoV-2 patients, this procedure was repeated for cohorts containing only COVID patients and only patients without any COVID infection. The model that was optimized and trained on the *NON-COVID* cohort was then evaluated on the the *COVID* cohort. This approach is detailed in Chapter 4.5.1.

4.5 Results

This section details the predictive performance of different models on the cohorts defined in Chapter 4.2. First, the model performance for the whole ASIC cohort is evaluated. Then we test if such a model could have been of use during the COVID-19 pandemic, by training a model on patients where no SARS-CoV-2 infection was present and evaluating the model on patients infected with the virus. Finally, the effect of over- or undersampling to reduce the impact of the inherent class imbalance, as described in Chapter 2.3, is shown.

Hyperparameter tuning and model training utilized the computing resources provided by the RWTH Aachen University, most notably server nodes equipped with two NVIDIA Tesla V100 GPUs each. Simulations were performed with computing resources granted by RWTH Aachen University under project rwth1547, as well as resources granted by the Joint Research Center for Computational Biomedicine.

The predictors utilize the data layout described in Chapter 4.2 and predict, based on the past 24 hours of data, whether a stark decline in Horowitz index, as defined in Chapter 4.3, will occur two to three days from the time of prediction. Thresholds for the model are chosen to maximize the sensitivity for the minority class.

Figure 4.11 shows the predictive performance of the model at different thresholds through both ROC and PR curves. The classfifier achieved a ROC-AUC of 0.90 and a AUC-PR of 0.54, at a prediction horizon of three days prior to the event. Additional metrics are presented in Table 4.6. Most notably the model operates at a specficity of 0.47 and a sensitivity of 0.99. A more balanced threshold prioritizing the macro averaged F_1-Score results in a specificity of 0.92 at a sensitivity of 0.65. The corresponding optimized

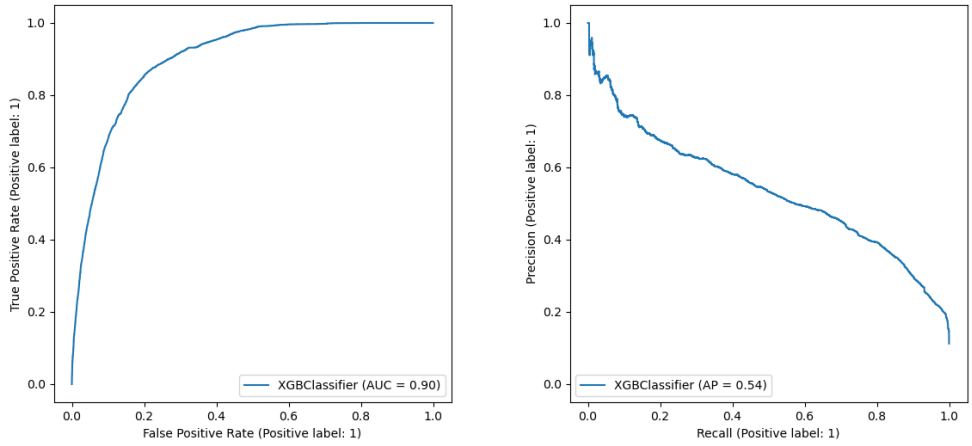


Figure 4.11: Receiver Operating Characteristic and Precision-Recall curves for the predictor on the ASIC cohort. The model operates at a prediction horizon of three days, using data from the post 24 hours as input.

F_1-Score is 0.74.

Table 4.6: Model performance

AUC-PR	ROC-AUC	F_1-Score (Macro Averaged)	Specificity	Sensitivity
0.54	0.90	0.48	0.47	0.99

The model was evaluated on 10% of the data available, with 90% of the data used for training. For training 662,424 samples were used and for testing 73,661. Due to the stratified splitting the event rate in both datasets was comparable at 11.237% and 11.239% each. The confusion matrix for the model is shown in Table 4.7.

Table 4.7: Confusion Matrix for the ASIC model

		Predicted	
		0	*1*
Actual	*0*	30,656	34,682
	1	73	8,200

Table 4.8 compares the developed model to the state of the art as introduced in Chapter 4.1. Note that the surrogate marker deployed in this work is a different prediction target than the adapted Berlin criteria deployed by other authors. We argue that a CDSS focussing on rapid loss in oxygenation is often clinically more relevant than a model

focussing on the first onset of ARDS. Figure 4.5 indicates some issues with ARDS onset, as defined by a PaO_2/FiO_2 ratio $\leq 300/200/100$ and a $PEEP \geq 5\ cmH_2O$. For the given example patient the $PEEP \geq 5$ criteria is fulfilled for almost the whole stay, predicting every single time the Horowitz index would drop below the 300/200/100 threshold would thus happen very often, while the clinical relevance would remain limited. The surrogate marker defined in Chapter 4.3 would only sound alarm five times in contrast to the 34 times the Berlin thing would.

Table 4.8: Comparison of our developed model with various other ARDS prediction models. Le et al. and Singhal et al., both use comparable ML pipelines but focus on a different prediction target.

Authors	Dataset	Model Type	Horizon	Target	Best ROC-AUC	Sensitivity	Specificity	Notes
Zaglam et al.	TARD	LDA	—	ARDS	—	0.91	0.87	Detect ARDS in Radiograph
Taoum et al.	MIMIC-II	Novelty	Up to 39h	Berlin	0.79	0.65	1.00	Small Cohort, HighRes data
Le et al.	MIMIC-III	XGBoost	Up to 48h	Berlin	0.90	0.67	0.85	Includes Radiology Reports
Singhal et al.	Multicenter	XGBoost	Up to 36h	Berlin	0.89	0.77	0.85	Focus on COVID patients
Polzin et al.	ASIC	XGBoost	Up to 72h	Novel	0.90	0.65	0.92	Our model

To further evaluate the model setup in this work, it was compared to the work by Le et al. [9] by adapting a similar prediction target of PaO_2/FiO_2 ratio $\leq 300/200/100$ and a $PEEP \geq 5\ cmH_2O$. The ASIC dataset does not include radiology reports, which were embedded in the model by Le et al. Further, Le et al. excluded patients with a Tracheostomy in the first 72 hours. The paper by Le et al. includes all ICU patients for their model, while the ASIC cohort only included patients mechanically ventilated for at least 24 hours. Le et al. provide results for a mechanically ventilated subcohort, with at least one hour of mechanical ventilation in a supplement to their publication though. All differences are listed in Table 4.9. Table 4.10 shows the 10-fold cross-validated ROC-AUC achieved at the different prediction horizons.

The state of the art in this field is notably heterogeneous, as patient cohorts and

prognostic features are typically tailored to the specific objectives of individual studies and analyses, making direct comparisons challenging. Furthermore, access to data from such studies is often restricted due to privacy or data governance considerations. Our decision to compare our work with that of Le et al. was motivated by the high degree of similarity between the two in several key aspects. However, as highlighted in Table 4.9, significant differences remain and must be carefully considered.

Table 4.9: Differences in data and model setup compared to Le et al.

Model	Dataset	Radiology Reports	MV	Age >= 18y	Total Count
Le et al.	MIMIC-III	yes, as Word2Vec	>=1h	yes	9133
DEA Model	ASIC	no	>=24h	yes	3676

Table 4.10: Comparison of predictive ROC-AUC achieved at different prediction horizons deploying our developed model to the ASIC dataset to the model developed by Le et al. to their MIMIC-III cohort.

Model	At Onset	12h	24h	48h
Le et al.	0.843	0.858	0.810	0.796
DEA Model	0.965	0.975	0.960	0.897

4.5.1 Generalization

To test the generalizability of the model with respect to different causes of ARDS the ASIC dataset was split in two sub-cohorts based on ICD Codes for the presence of SARS-CoV-2, as described in Chapter 4.2. Table 4.2 describes the two cohorts in more detail. The *COVID* cohort contains significantly less encounters, with 296 visits to the ICU recorded. The *NON-COVID* cohort on the other hand encompasses 3,380 encounters, being more than eleven times larger than the *COVID* cohort. Additionally, ARDS is much more prevalent in the *COVID* cohort, with 84.1% developing ARDS, in contrast to 17.2% for the *NON-COVID* cohort. The distribution of ARDS severity also differs between the cohorts as shown in Table 4.11. In the *COVID* cohort significantly less mild ARDS and moderate ARDS is present, but severe ARDS is increased.

Table 4.11: Comparison of relative ARDS severity percentages.

Cohort	Mild ARDS	Moderate ARDS	Severe ARDS
FULL	6%	41%	53%
NON-COVID	8%	45%	47%
COVID	2%	32%	66%

Table 4.12 provides an overview of the respective best-performing models. The *NON-COVID* model performs better than the *COVID* model concerning AUC-PR, indicating a better overall performance at different thresholds of precision-recall. The *COVID* model, at a sensitivity of 0.99, achieves a specificity 0.07 lower than the *NON-COVID* model. The F_1-Score score achieved by the *COVID* model is also 0.10 lower than that of the *NON-COVID* model.

Table 4.12: Comparison of the final models on *COVID* and *NON-COVID* cohorts, respectively.

Cohort	AUC-PR	ROC-AUC	F_1-Score (Macro Averaged)	Specificity	Sensitivity
COVID	0.50	0.90	0.36	0.37	0.99
NON-COVID	0.58	0.91	0.46	0.44	0.99

The respective confusion matrices for both predictors are presented in Table 4.13 and Table 4.14. Tuned for maximum sensitivity on the rapid loss of oxygenation, the *NON-COVID* model would raise a total of 39,211 alarms, with 81% of them being false. The model would in turn though only miss 1.6% of the 7,596 events.

Table 4.13: CM for *NON-COVID* predictor

		Predicted	
		0	*1*
Actual	*0*	23,473	31,736
	1	121	7,475

Table 4.14: CM for *COVID* predictor

		Predicted	
		0	*1*
Actual	*0*	3,402	5,706
	1	2	666

The ROC and PR curves for both models can be found in the Appendix.

To evaluate how well the model can generalize from a population not infected with SARS-CoV-2 to a population infected by the virus, a model denoted *transfer* is

trained on the *NON-COVID* cohort and then used to predict on the *COVID* cohort. This model has significantly more training data available. It improves on the area under the precision-recall curve by 0.02 and achieves a 0.15 higher F_1-Score when predicting the *COVID* cohort in contrast to the model trained natively on the cohort. Comparing the CM for the transfer model in Table 4.16, we see 58 more false negatives and 2249 fewer false positives for the model. Figure 4.12 shows the difference in the respective precision-recall curves, indicating the *transfer* model offers a better tradeoff between precision and recall in most cases.

Table 4.15: Comparison of Predictors. *transfer* refers to a model that is trained on the *NON-COVID* cohort and predicts the *COVID* cohort. The model outperforms the native *NON-COVID* predictor at a higher sensitivity with a comparable specificity.

Cohort	AUC-PR	ROC-AUC	F_1-Score (Macro Averaged)	Specificity	Sensitivity
COVID	0.50	0.90	0.36	0.37	0.99
NON-COVID	0.58	0.91	0.46	0.44	0.99
transfer	*0.52*	*0.90*	*0.51*	*0.62*	*0.91*

Table 4.16: Confusion Matrix for the transfer model.

		Predicted	
		0	*1*
Actual	*0*	5,651	3,457
	1	60	608

4.5.2 Random Over- and Under-Sampling

The given data is imbalanced, with an overall event rate of 11.2% percent. This type of imbalance problem often exists in the medical domain [192] and can lead to various problems [193]. Due to the increased prior probability, learners will often overestimate the majority group in such datasets [194]. Rare minority samples may be treated as noise or noise may be incorrectly classified.

This topic was already discussed to some extent in Chapter 2.3, with a focus on evaluation metrics. Two additional techniques for dealing with such imbalanced data were explored and compared: Random Over Sampling (ROS) and Random Under Sampling (RUS) [195]. In undersampling, data from the majority class - timepoints where no drop in Horowitz index occurs - is removed from the training data. This balances the dataset, alleviating the abovementioned issues with imbalanced data. Undersampling to equal class

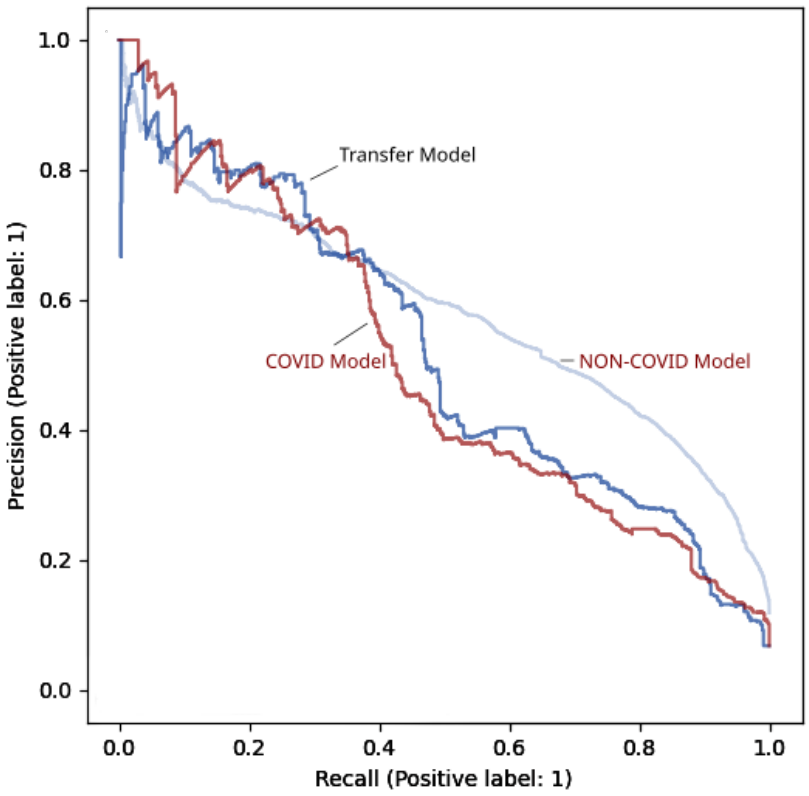


Figure 4.12: Comparison of the precision-recall curves for all three evaluated predictors. The generalizing model performs worse in general than the *NON-COVID* model on its native cohort but outperforms the *COVID* model by 0.13 when applied to patients infected with SARS-CoV-2.

sizes reduces the size of the majority class to a fraction of its original size. Oversampling creates artificial data from the minority class. Either through duplicating existing data or creating synthetic data from the distributions of the minority class.

Table 4.17 shows the different event rates and dataset sizes depending on the sampling strategy used, with the untouched data as a baseline.

Table 4.17: Overview of data available for training and event rates for different sampling strategies.

Cohort	Sampling Strategy	Training Data Size	Event Rate
COVID	Baseline	100,137	6.83%
	Oversampling	195,624	50%
	Undersampling	4650	50%
NON-COVID	Baseline	709,301	11.91%
	Oversampling	1,374,356	50%
	Undersampling	44,246	50%

The results collected in Table 4.18 show the results for over- and undersampling the data prior to training the models. For the *COVID* cohort, both strategies are able to improve the overall model performance, as described by the AUC-PR and ROC-AUC. The best achievable F_1-Score decreases though. For the much larger *NON-COVID* cohort no such improvement can be observed and the F_1-Score decreases as well.

Table 4.18: Comparison of results for different sampling strategies.

Cohort	Sampling Strategy	AUC-PR	ROC-AUC	F_1-Score (macro averaged)
COVID	Baseline	0.41	0.94	0.74
	Oversampling	0.45	0.95	0.61
	Undersampling	0.44	0.95	0.58
NON-COVID	Baseline	0.57	0.95	0.69
	Oversampling	0.57	0.95	0.54
	Undersampling	0.54	0.95	0.54

4.6 Discussion

This chapter introduced a prediction model for the rapid loss of oxygenation in intensive care patients. The definition of *rapid loss* is leaning on the thresholds defined by the Berlin definition and was fine-tuned in collaboration with medical experts. For patients in a less severe state with respect to oxygenation, i.e. patients that currently have a high Horowitz index, the drop in this parameter needs to be larger than in patients with already low oxygenation. Further, the definition always requires a drop of at least 50 in between the two windows. Any slower change is not considered a *rapid* decline and should be handled by a knowledge-based CDSS, monitoring the Berlin criteria, akin to the ASIC App, developed as part of this research project. It is noteworthy, that the surrogate marker defined in Chapter 4.3 does not include the need for a $PEEP \geq 5$ as included in the Berlin definition. While crucial for the definition of ARDS, an alarm system in the ICU, with the broader prediction target of loss of oxygenation, will be clinically more relevant. Even if a patient has a PEEP of 4 and a decrease from a Horowitz index of 300 to one of 80 would not indicate ARDS, this still definitely should raise an alarm and a medical professional should be notified.

The prediction target that is proposed is both more ambitious and more clinically relevant than a predictor for ARDS based on $PEEP \geq 5$ and Horowitz index ≤ 300. It still captures patients that develop a rapid onset ARDS, as all these patients will see a stark decrease in oxygenation, which the developed predictor is focussed on. Additionally, the developed CDSS alarm system will take into account various pathologies causing a loss of oxygenation. The developed predictor performs well, with performance metrics comparable to, or outperforming, similar approaches. The model is tuned to be very sensitive to the minority class of loss-of-oxygenation events. Only 1.6% of these events are missed by the model, indicating good predictive performance. The prediction horizon of up to three days allows for early intervention and could be used to screen potential candidates for study participation, which is often an issue in ARDS research [165].

The pipeline developed outperforms the approach by Le et al., as seen in table Table 4.10, for a similar prediction target. This mostly reinforces that the developed model performs well, as there are various significant differences between the two applications.

Two completely distinct cohorts are used, with our model being focused on patients at least mechanically ventilated for 24 hours from the ASIC dataset, while their model is screening all ICU patients in MIMIC-III. Further, Le et al. included radiology reports, both to assert the presence of ARDS if reports mention bilateral opacities or infiltrates, as required by the Berlin criteria, and as additional input features to their model. Many implementation details may also differ, as the source code for their implementation is not publicly available, and various details of the implementation are thus not available. For example, default parameters of the XGBoost framework, not explicitly mentioned in their publication, could have changed in between versions of the software, implicitly affecting the parameterization of both models respectively. Nonetheless, the comparison indicates that the model developed is capable of performing adequately on a similar prediction task.

Considering the dangers of "alarm fatigue", as introduced in Chapter 1, it is necessary to judge models intended for use in the ICU carefully, balancing the risk for patients in critical states without an alarm on the one side and the alarm fatigue on the other side. The best-performing model in this retrospective study would have sounded 42,882 alarms. 8,200 (19%) of these alarms would have been relevant. A total of 73 (0.9%) events would have been missed. The risk of missing an event of interest, in contrast to the issues caused by alarm fatigue, is highly individual and can further depend on every patient's current state. In the example above the model is tuned to be as sensitive as possible. For a patient that is in a less dangerous situation, it might be suitable to lower the sensitivity of the model, reducing the rate of false alarms in the process. Such a risk-adaptive strategy, however, raises ethical concerns and has to be developed and implemented in close integration of ethical consensus. Direct feedback from medical experts, as well as the comparison to other ARDS predictors, indicates that the model developed in this work would provide value to the ICU. While it is still generating many false alarms, it also shows good predictive performance for the targeted events. As ARDS is associated with high mortality and morbidity, a predictive early warning CDSS, even if it produces some false alarms, could improve patient outcomes significantly. In contrast to some alarm systems already in use in the ICU, with reported false alarm rates of up to 88% [52], the 81% false alarms generated by our model might be feasible, considering a focus on high sensitivity. Developing models that are more precise should remain a strong

research focus though, as every new model introduced to the ICU, even if performing better than existing ones, will contribute to the noise and the increasing challenges of alarm fatigue.

Considering the increasing risk of new pandemics [196] [197] [198], being well prepared for such outbreaks is becoming more and more relevant. One strategy to prepare for pandemics is the development of prediction models that target syndromes and conditions that are shared by different infections and diseases. Examples of such prediction targets include ARDS, Sepsis, acute kidney injury, liver failure, or multi-organ dysfunction. This work shows that a model trained on a non-pandemic population of patients who develop ARDS can be deployed to generate alarms for a pandemic population. Research on the clinical features observed in COVID-19 patients who develop an ARDS suggests that the syndrome developed in those patients does not differ significantly from those not infected by SARS-CoV-2 [199]. The applicability of our developed model trained on NON-COVID patients to the COVID population supports these results.

Hypothesizing this to be possible for other conditions and different pathogenies could prove to be a valuable strategy in pandemic preparedness. As discussed previously, the balance between per-patient risk and alarm fatigue is highly individual. While a model may not be suitable for day-to-day ICU use, this may change in a pandemic situation. Considering the high lethality of some SARS-CoV-2 mutations a prediction model that produces many false alarms might be helpful, even if it produces a lot of noise, if an early detection could lead to a prevention of a rapid deterioration in the patient's state. Several factors limit this hypothesis; only data from a single hospital has been used. This limits the generalizability and introduces a risk of hospital bias. Further, the hypothesis has only been tested for COVID-19.

More generally, this work is limited in various ways. The alarm model proposed in this study has only been evaluated retrospectively. The data for training could have been processed in different ways, and more strategies to both describe and format the data could have been explored. Feature engineering on the data windows in particular was only explored briefly, and this research focused on XGBoost models early on, not exploring alternatives in depth. One potential source of bias in this context stems from

the reliance on laboratory tests to determine PaO2. Due to the required BGA the sampling rate for PAO2 is highly irregular. It is further influenced by the medical experts, as a doctor who is already anticipating an ARDS to develop is likely to request more regular BGA testing to be done to confirm his suspicion.

The results indicate that the overall model performance for the smaller *COVID* cohort could be increased by both over- and undersampling. For the *NON-COVID* cohort no such increase was found. For both cohorts the highest achievable F_1-Score was decreased though, indicating a lower peak performance for our metric of choice, when data was resampled. This could be related to the simplicity of sampling strategies. More sophisticated sampling approaches, such as SMOTE [200] or SEB-XGB [201], have been shown to improve performance in imbalanced learning tasks for XGBoost models [202] [203] [204]. The decrease in performance when working with with undersampling for XGBoost is further in line with other research [205] [206], suggesting that random under-sampling deteriorates XGBoosts performance in general.

Additionally, XGBoost has various parameters that can impact the training behavior and help with imbalanced learning tasks. They were explicitly explored in hyperparameter tuning to help with the imbalance. For example, the classes can be assigned different weights, which are used to scale the gradients during training, making a model overcompensate for the lack of samples in the minority class. The recommended value for imbalanced learning tasks is:

$$scale\_pos\_weight = \frac{\text{total negative examples}}{\text{total positive examples}}$$

If the risk of missing an upcoming drop in Horowitz index for a patient outweighs the issues caused by an increased rate of false alarms, exaggerating the minority class (decrease in Horowitz index) during training can lead to better performance at the cost of more false alarms. As this scenario could be desirable in a hospital setting, we also analyzed the impact of higher weights for the minority class. The full list of values for *scale_pos_weight* explored is detailed in the Appendix. Additional parameters of the XGBoost model that were explicitly included are *min_child_weight* and *max_delta_step*,

which define how conservatively the model behaves to ensure the minority class is not an oversight during training.

Chapter 5

Conclusion

This work makes two scientific contributions. First, a novel framework for the development of ML applications for medical timeseries data is presented. This framework, the Diagnostic Expert Advisor, provides various crucial features to researchers. Most notably, it enables a straightforward integration of HPC, an interactive workflow through visualization, explorative analysis, and a common codebase to reuse and more easily share research code. Secondly, a predictive model for intensive care patients is presented. This model predicts a rapid decline in the lungs capacity to oxygenate the bloodstream from routinely charted ICU data. It achieves high predictive performance, comparable to similar approaches found in literature, while operating at a larger prediction horizon of up to three days. This model was developed in the context of ARDS, and the prediction target encompasses possible onset times of rapidly developing ARDS. It is strongly inspired and shaped by clinical relevance and usability though, focusing on oxygenation in general, rather than the specific thresholds defined by the Berlin definition for ARDS diagnosis, covering other possible causes of a decrease in oxygenation as well.

The first (1) contribution of this work is the Diagnostic Expert Advisor. A platform for model development on timeseries data. Medical data is often sparse and heterogeneous, particularly time series data from the ICU, which frequently exhibits these characteristics. Working with such data requires additional care and often increased

computational resources. In many cases, the use of HPC becomes essential, especially when applying machine learning methods. While some areas of medical research, such as imaging, have a robust set of tools and platforms for integrating HPC, to the best of our knowledge no open-source platforms dedicated to the development of time series models on medical data exist. The DEA addresses this gap by providing an open-source, freely available platform to advance machine learning research in this domain of medical research. It fills a crucial gap in research software at the intersection of medicine and ML, enabling and supporting cooperation and collaboration between medical experts and researchers The DEA is designed to the FAIR principles [207]. This concept was originally developed with respect to data management and promotes the goal of open science by making data findable, accessible, interoperable, and reusable [208]. A proposed adaptation to software reinforces research software as a fundamental and vital part of the research landscape and tackles corresponding challenges in "discoverability, productivity, reproducibility, and sustainability" [209].

It further promotes the use of HPC resources by providing an intuitive way to parallelize software. This makes HPC more accessible to medical researchers with limited computer science training. As a result, HPC adoption in the medical field is encouraged. Underlying technologies deployed in the DEA are in widespread use in the ML community, enabling the use of established tools for complex data processing and modeling. Interoperability with existing frameworks allows for a high degree of freedom in design and development. By providing a patient-centric interface, similar to systems deployed in hospitals, the DEA encourages a standardized and intuitive research workflow in the medical context. The built-in focus on visualization and exploration promotes discussion and exchange in cooperative research projects. Model development with the DEA is focused on the intensive care environment, where heterogeneous data is available at high resolution and syndromes and complications can develop rapidly, leading to swiftly deteriorating patient states. Integrating these high-dimensional data streams with a framework that enables the combination of various modeling approaches provides a platform allowing for the fast and streamlined development of ML predictors for dangerous syndromes and conditions, such as ARDS, Sepsis, or acute kidney injury.

Another essential scientific impact of the DEA platform is its value concerning replication. Replication describes the ability to confirm results by repeating the steps outlined by the original researchers. It is considered by some to be "the cornerstone of science" [210] [211]. As software is at the core of many research areas, it is essential for replication. While research software could allow for significantly easier replication of study results, it is often a significant hindrance. In many cases, the code that was used is not made available. For example, a recent study found that up to 24% of papers are now shared with code available in image processing. While this percentage has more than tripled since 2004, it is still barely a fourth of all publications analyzed. The authors found that research with code published was cited twice as often, but the overall incentive to publish research software still seems lacking [114]. In another example, Odd Erik Gundersen analyzed the code availability from 400 algorithms presented at two top AI conferences, finding only 6% of the presenters sharing code [212]. Researchers in Germany identified various challenges in this regard [213]. For example, lack of individual benefits, incentive systems, awareness, expertise, or legal-, as well as funding issues. They summarize these as a lack of sufficient incentivization and point out that the practice of creating such software does not align well with the publish-or-perish culture. Issues in replication stem from many reasons, other factors being, for example, publication bias [214], the aforementioned pressure to publish [215] [216], or confirmation bias [217].

The FAIR principle has been established in research data to enhance the data's usability and impact. The guidelines defined for research data have been adapted to research software, with further extensions being discussed and developed in the research community [209], further integrating best practices of software development [218]. In addition to being published and accessible; good research software should be discoverable, licensed under a suitable open source license, and have clearly defined contribution, governance, and communication processes [218]. The DEA itself fulfills all these criteria and further encourages researchers using it to adhere to the same principles. Taking advantage of these design paradigms, the developed software platform is currently being utilized in the National High Performance Computing Center for Computational Engineering Science (NHR4CES) Simulation and Data Lab (SDL) Digital Patient. The platform is used in training courses and seminars to encourage clinicians and researchers in biomedical

research to pursue the utilization of HPC in their work. Through the platform's modular and patient-centric design, it can also serve as a foundation for further development toward digital twin (DT) and virtual patient (VP) modeling. By increasing personalization of the developed models, the DEA can be fundamentally extended to encompass not just data-driven ML models but also, for example, mechanistic ones, informed by physical and medical knowledge and custom fit to each patient. In this context, the DEA is laying a foundation for digital twin modeling in the European Virtual Human Twin (EDITH) project, functioning as a digital twin modeling platform demonstrator.

This work's second (2) contribution is a prediction model for intensive care patients that predicts an upcoming rapid decline in the FiO_2/PaO_2 ratio. This ratio expresses the lungs ability to oxygenate the bloodstream and is considered an indicator of lung health. A rapid loss could be caused, for example, by ARDS. Notably, no matter which underlying condition is driving this change, such a stark decrease is clinically relevant, and an early alarm can alert medical experts of a patient's deteriorating state. All patients that develop rapid onset ARDS will be identified by the model, as the surrogate marker defined for ARDS is more extensive, but covers rapid onset ARDS. Comparing the predictor developed in this work to the current literature we achieve a comparable ROC-AUC, Sensitivity and Specificity, at a larger prediction horizon, while the prediction target for the model we developed is more ambitious and clinically relevant. The predictor developed misses less than 1% of the events of interest. At this threshold our model has a false alarm rate as high as 81%, comparable to systems actively in use in the ICU, which can exhibit false alarm rates of up to 88% [52]. Hence, our model improves the true positive rate by almost 50%, nevertheless the false alarm rate remains high.

Due to the unique ASIC cohort, the model was trained and tested on, and the novel prediction target, a direct comparison to other ARDS predictors does not pose an adequate comparison. We thus compared the developed machinery to other models found in literature by adopting a prediction target similar to that deployed by Le et al. and Singhal et al.. This definition closely resembles the Berlin definition for the diagnosis of ARDS and is focused on a PaO_2/FiO_2 ratio $\leq 300/200/100$ and a $PEEP \geq 5\ cmH_2O$. We compared the results achieved on the ASIC dataset with this prediction target to those

achieved on MIMIC-III by Le et al.. The cohorts originate from different hospitals and patients included differ in the required length of mechanical ventilation. Le et al. further incorporated radiology reports for both diagnosis of ARDS and as input features for their model. The comparison shows that the developed model pipeline works well, either outperforming the predictor by Le et al. or operating at a similar performance level.

The retrospective applicability of the model in the COVID-19 pandemic was explored by applying a model trained exclusively on patients without SARS-CoV-2 infection to those who got infected. In contrast to a dedicated model, trained and tested on the infected cohort, the performance was significantly increased. This shows, that a model that was trained and setup before the pandemic spread, could have helped predict ARDS in patients infected with the virus. Research indicates that the pathology of ARDS in COVID patients does not differ significantly from those in patients without the infection. The results of this experiment support these findings. The development of general predictors for life-threatening conditions and syndromes that can arise from different diseases could prove vital, both in general ICU CDSS, and in pandemic preparedness. Our retrospective analysis showed, that an established model for the prediction of a sudden loss in oxygenation, would have been capable of predicting upcoming oxygenation issues in COVID-19 patients during the pandemic.

In conclusion, the work presented in this thesis makes two contributions to the current research landscape. The overarching goal is to develop a predictor that could improve the outcome of patients developing ARDS in the ICU. To achieve this goal, a platform for the development of ML model on heterogeneous medical timeseries data was created. This platform was then used to set up a predictor for stark decreases of oxygenation in intensive care patients, inspired by ARDS. This model proposes the use of a novel prediction target, focusing on deterioration of patient health, instead of fixed thresholds used to diagnose severity of ARDS and achieves high predictive performance using this target, as well as using the in literature more commonly used thresholds based on the Berlin definition. The development of ML models on the vast and constantly increasing data available in hospitals, especially in intensive care, will certainly become increasingly relevant in the future, and the application of the platform in the context of

ARDS shows its viability. Academic research sometimes diverges from clinical needs and usability, especially when research is conducted in an "ivory tower" and contact between medical experts and researchers can be limited. Many consortia promote and emphasize interdisciplinary collaboration to ensure the applicability of research outcomes in the clinics and a focus on research that can significantly impact daily practice, or focuses on areas the medical experts agree on being most relevant. The focus on oxygenation for the proposed prediction model originates from such a interdisciplinary collaboration. We consider the focus on rapid loss of oxygenation in a patient significantly more clinically relevant then a model predicting ARDS crossing the thresholds defined in the Berlin definition by a PaO_2/FiO_2 ratio ≤ 300 and a $PEEP \geq 5\ cmH_2O$. This work suggests, that a CDSS focussing on these thresholds, while important for patients with a slowly developing ARDS, falls short for patients with faster trajectories, generating many alarms, with limited significance. Our proposed prediction target on the other hand focuses on fewer, more relevant events.

While the performance achieved by the predictor we developed is promising, multiple limitations still need to be pointed out. Most notably the prediction model was developed and evaluated only retrospectively. Thus, it is not capable of capturing the complex interactions of medical experts with such a model. Further it does not incorporate any interaction with treatment options for ARDS, and other causes of rapid oxygenation loss. This research is further focussed on a single hospital, the University Hospital Aachen, from the ASIC cohort, as suitable data from other hospitals was not available at the time of writing. Most of this work was conducted while data was still being collected from other locations participating in the ASIC study, and the pipeline for the analysis of potential hospital biases was still in development [14]. The results thus may not transfer well to other wards or hospitals, especially if not focused on intensive care.

While the proposed model demonstrates strong predictive performance, several obstacles prevent its direct implementation in the ICU. Conceptually, particularly in retrospective analyses, the sparse and heterogeneous nature of the data poses significant challenges. For example, missing data is often not random but may reflect deliberate decisions by the treating physician to order specific tests based on their judgment. As a

result, the data is inherently shaped by what the clinician deems important, creating a causality dilemma similar to the classic 'chicken or egg' problem. Novel approaches and careful evaluation are necessary, especially as we move toward prospective analysis, to ensure that machine learning models are not merely "looking over the shoulders" of experts [219]. On a practical level the high false alarm rate indicates a need to further increase model performance. Future work should focus on improving model performance, as to not contribute to alarm fatigue more than necessary. The DEA as a platform bears the potential to extend the presented prediction model further. Identifying sub-cohorts which exhibit similar behavior can allow the application of more specific and fine-tuned models, for example a dedicated predictor for ARDS caused by pneumonia, and a different predictor for sepsis-induced ARDS. Preliminary results have indicated that such dedicated models can outperform a generic model trained on the pooled data, but identifying relevant sub-cohorts remains challenging. Another strategy to further improve predictions lies in the increasing personalization of such models to integrate additional valuable information. On the one hand, mechanistic models can be fit to individual patients. On the other hand, ML approaches can be deployed on an individual level, for example individually trained models for parameters such as heart rate, or FiO_2. Perpetuating along this path we deem virtual patients and digital twin modeling crucial research areas in the future of personalized prediction of ARDS.

Appendix

List of parameters present for the FULL cohort

		Missing	Overall
n			7237099
24h-Bilanz (Fluessigkeiten-Einfuhr vs -Ausfuhr), mean (SD)		7170350	40.2 (1484.0)
AF, mean (SD)		1706653	19.4 (6.7)
AF spontan, mean (SD)		4467588	9.9 (10.9)
Albumin, mean (SD)		7204851	473.7 (2009.8)
Amylase, mean (SD)		7234144	117.1 (563.5)
BE arteriell, mean (SD)		6677001	2.4 (4.4)
Bicarbonat arteriell, mean (SD)		6675324	27.1 (4.5)
Bilirubin ges., mean (SD)		7168424	26.2 (57.8)
BMI, n (%)	1	0	625 (17.0)
	2		290 (7.9)
	3		205 (5.6)
	L		97 (2.6)
	M		898 (24.4)
	P		1551 (42.2)
	nan		10 (0.3)
BNP, mean (SD)		7230479	5943.4 (11892.9)

		Missing	Overall
CK, mean (SD)		7139656	241.6 (1209.0)
CK-MB, mean (SD)		7222405	50.4 (92.4)
clusterAlter, n (%)	70-79	0	922 (25.1)
	80-130		487 (13.2)
	<70		2267 (61.7)
clusterGeschlecht, n (%)	M	0	2401 (65.3)
	W		1275 (34.7)
clusterKoerpergewicht, n (%)	65-75	0	908 (24.7)
	76-250		2277 (61.9)
	<65		491 (13.4)
clusterKoerpergroesse, n (%)	180-185	0	836 (22.7)
	<180		2584 (70.3)
	>185		256 (7.0)
Compliance, mean (SD)		5710987	56.7 (40.2)
CRP, mean (SD)		7192457	1061.7 (785.4)
D-Dimere, mean (SD)		7222318	17791.4 (40591.2)
DAP, mean (SD)		1618104	59.6 (13.0)
deltaP, mean (SD)		5322926	14.0 (4.9)
DPAP, mean (SD)		7000469	22.6 (7.6)
ECMO, mean (SD)		986554	0.0 (0.0)
etCO2, mean (SD)		7196757	41.5 (11.0)
Extrakorporaler Blutfluss, mean (SD)		7232829	3.9 (1.0)
Extrakorporaler Gasfluss (O2), mean (SD)		7226880	6.2 (8.2)
FiO2, mean (SD)		4421204	42.4 (15.3)
Gaszusammensetzung (%O2), mean (SD)		7226541	85.2 (23.0)
GOT, mean (SD)		7167486	141.7 (664.6)
GPT, mean (SD)		7168636	93.7 (259.7)
Haematokrit, mean (SD)		6598392	28.9 (4.9)

	Missing	Overall
Haemoglobin, mean (SD)	6668922	5.9 (1.0)
Harnstoff, mean (SD)	7164988	10.8 (7.1)
HF, mean (SD)	1142530	86.7 (18.1)
Horowitz-Quotient (ohne Temp-Korrektur), mean (SD)	6574613	313.0 (210.9)
I:E, mean (SD)	5758499	1.9 (1.3)
IL-6, mean (SD)	7225503	1704.6 (26591.4)
individuelles Tidalvolumen pro kg idealem Koerpergewicht, mean (SD)	6914897	402.5 (62.8)
Inhalatives NO, mean (SD)	7171699	16.6 (8.0)
INR, mean (SD)	7142999	1.2 (0.5)
iSOFA.HKL, mean (SD)	5883182	1.6 (1.6)
iSOFA.iSOFA Gesamt, mean (SD)	5883182	5.7 (4.4)
iSOFA.Leber, mean (SD)	5883182	0.4 (1.0)
iSOFA.Lunge, mean (SD)	5883182	1.3 (1.2)
iSOFA.Niere, mean (SD)	5883182	1.2 (1.6)
iSOFA.Thrombo, mean (SD)	5883182	0.5 (0.9)
iSOFA.ZNS, mean (SD)	5883182	0.7 (1.3)
Koerperkerntemperatur, mean (SD)	2189529	36.7 (2.6)
Kreatinin, mean (SD)	7165711	101.2 (86.1)
Lagerungstherapie, mean (SD)	986554	0.0 (0.0)
Laktat arteriell, mean (SD)	6674857	1.4 (1.6)
LDH, mean (SD)	7198996	458.2 (786.0)
Leukozyten, mean (SD)	7155377	12.1 (9.7)
Lipase, mean (SD)	7212386	84.1 (259.6)
Lymphozyten absolut, mean (SD)	7233249	1.3 (0.7)
Lymphozyten prozentual, mean (SD)	7233246	0.1 (0.1)
MAP, mean (SD)	1578851	79.9 (15.8)
MPAP, mean (SD)	6991044	29.3 (9.4)

		Missing	Overall
P EI, mean (SD)		4475934	20.4 (7.9)
paCO2 (ohne Temp-Korrektur), mean (SD)		6865579	42.7 (9.9)
paO2 (ohne Temp-Korrektur), mean (SD)		6820377	89.6 (33.9)
PCT, mean (SD)		7179789	2.8 (9.4)
PCWP, mean (SD)		7233458	16.7 (6.0)
PEEP, mean (SD)		4800589	7.3 (2.7)
PEEP eingestellt, mean (SD)		6939314	8.1 (2.3)
Phase, n (%)	0	0	1081 (29.4)
	1		358 (9.7)
	2		2237 (60.9)
pTT, mean (SD)		7140034	38.5 (15.0)
SaO2, mean (SD)		6675986	92.7 (10.7)
SAP, mean (SD)		1615759	123.6 (24.4)
SOFA.Punkte Blut, mean (SD)		7173507	0.5 (0.9)
SOFA.Punkte Leber, mean (SD)		7178121	0.4 (1.0)
SOFA.Punkte Lunge, mean (SD)		7174787	2.0 (0.9)
SOFA.Punkte Lunge (cal.), mean (SD)		7223560	1.7 (0.9)
SOFA.Punkte Niere, mean (SD)		7170929	3.6 (0.8)
SOFA.Punkte ZNS, mean (SD)		7171152	0.9 (1.5)
SOFA.SOFA Gesamt, mean (SD)		7170929	7.9 (3.2)
SPAP, mean (SD)		7000377	40.9 (13.1)
SpO2, mean (SD)		1839959	96.3 (3.8)
SVRI, mean (SD)		7235593	1492.4 (691.4)
SzvO2, mean (SD)		7202839	73.0 (16.2)
Thrombozyten, mean (SD)		7155708	230.0 (144.6)
Troponin, mean (SD)		7227776	0.7 (2.9)
ZVD, mean (SD)		5705859	11.7 (8.8)

List of parameters present for the COVID cohort

		Missing	Overall
n			893452
24h-Bilanz (Fluessigkeiten-Einfuhr vs -Ausfuhr), mean (SD)		885565	-48.3 (1246.9)
AF, mean (SD)		194207	20.1 (8.1)
AF spontan, mean (SD)		502667	7.4 (10.3)
Albumin, mean (SD)		887424	401.2 (611.7)
Amylase, mean (SD)		893157	58.2 (61.2)
BE arteriell, mean (SD)		814440	3.3 (4.9)
Bicarbonat arteriell, mean (SD)		814151	28.5 (5.0)
Bilirubin ges., mean (SD)		883775	28.3 (61.7)
BMI, n (%)	1	0	61 (20.5)
	2		35 (11.8)
	3		34 (11.4)
	L		3 (1.0)
	M		38 (12.8)
	P		125 (42.1)
	nan		1 (0.3)
BNP, mean (SD)		890259	4158.6 (9473.2)
CK, mean (SD)		880011	189.1 (1174.6)
CK-MB, mean (SD)		891192	28.9 (29.3)
clusterAlter, n (%)	70-79	0	59 (19.9)
	80-130		16 (5.4)
	<70		222 (74.7)
clusterGeschlecht, n (%)	M	0	205 (69.0)
	W		92 (31.0)
clusterKoerpergewicht, n (%)	65-75	0	42 (14.1)
	76-250		235 (79.1)
	<65		20 (6.7)

		Missing	Overall
clusterKoerpergroesse, n (%)	180-185	0	85 (28.6)
	<180		194 (65.3)
	>185		18 (6.1)
Compliance, mean (SD)		677226	43.6 (39.0)
CRP, mean (SD)		886845	1043.6 (821.8)
D-Dimere, mean (SD)		886318	17425.0 (34186.8)
DAP, mean (SD)		193379	59.5 (11.8)
deltaP, mean (SD)		499431	16.0 (5.4)
DPAP, mean (SD)		790964	23.0 (7.2)
ECMO, mean (SD)		147013	0.0 (0.0)
etCO2, mean (SD)		888059	44.7 (11.0)
Extrakorporaler Blutfluss, mean (SD)		892274	3.7 (0.9)
Extrakorporaler Gasfluss (O2), mean (SD)		889792	6.4 (3.1)
FiO2, mean (SD)		424295	47.1 (16.8)
Gaszusammensetzung (%O2), mean (SD)		890446	93.2 (15.4)
GOT, mean (SD)		883623	115.1 (547.6)
GPT, mean (SD)		883707	85.9 (186.9)
Haematokrit, mean (SD)		803146	30.5 (5.0)
Haemoglobin, mean (SD)		813473	6.2 (1.0)
Harnstoff, mean (SD)		885093	12.7 (7.3)
HF, mean (SD)		163094	92.2 (18.6)
Horowitz-Quotient (ohne Temp-Korrektur), mean (SD)		811994	267.3 (219.5)
I:E, mean (SD)		603609	1.6 (1.3)
IL-6, mean (SD)		889160	781.8 (11290.3)
individuelles Tidalvolumen pro kg idealem Koerpergewicht, mean (SD)		830936	406.2 (60.8)

		Missing	Overall
Inhalatives NO, mean (SD)		858610	17.0 (6.9)
INR, mean (SD)		880275	1.2 (0.3)
iSOFA.HKL, mean (SD)		708987	1.8 (1.6)
iSOFA.iSOFA Gesamt, mean (SD)		708987	7.1 (5.0)
iSOFA.Leber, mean (SD)		708987	0.4 (0.9)
iSOFA.Lunge, mean (SD)		708987	1.9 (1.3)
iSOFA.Niere, mean (SD)		708987	1.4 (1.8)
iSOFA.Thrombo, mean (SD)		708987	0.5 (0.8)
iSOFA.ZNS, mean (SD)		708987	1.2 (1.7)
Koerperkerntemperatur, mean (SD)		284477	36.8 (2.6)
Kreatinin, mean (SD)		885143	91.8 (80.0)
Lagerungstherapie, mean (SD)		147013	0.0 (0.1)
Laktat arteriell, mean (SD)		814247	1.3 (1.1)
LDH, mean (SD)		886745	512.5 (696.3)
Leukozyten, mean (SD)		881710	11.8 (6.1)
Lipase, mean (SD)		887612	92.9 (158.6)
Lymphozyten absolut, mean (SD)		891906	1.3 (0.7)
Lymphozyten prozentual, mean (SD)		891906	0.1 (0.1)
MAP, mean (SD)		189734	78.3 (14.4)
MPAP, mean (SD)		787464	30.1 (8.8)
P EI, mean (SD)		399542	23.4 (8.1)
paCO2 (ohne Temp-Korrektur), mean (SD)		835329	47.1 (12.0)
paO2 (ohne Temp-Korrektur), mean (SD)		828044	83.4 (34.7)
PCT, mean (SD)		885838	3.1 (8.2)
PCWP, mean (SD)		892100	15.6 (5.6)
PEEP, mean (SD)		491606	8.9 (2.7)
PEEP eingestellt, mean (SD)		854642	9.4 (2.8)
Phase, n (%)	0	0	35 (11.8)
	1		36 (12.1)
	2		226 (76.1)

	Missing	Overall
pTT, mean (SD)	879656	39.7 (13.4)
SaO2, mean (SD)	813933	93.1 (7.7)
SAP, mean (SD)	193114	118.8 (21.5)
SOFA.Punkte Blut, mean (SD)	885859	0.5 (0.8)
SOFA.Punkte Leber, mean (SD)	885935	0.4 (0.9)
SOFA.Punkte Lunge, mean (SD)	885836	2.3 (1.0)
SOFA.Punkte Lunge (cal.), mean (SD)	892800	1.4 (0.8)
SOFA.Punkte Niere, mean (SD)	885612	3.7 (0.8)
SOFA.Punkte ZNS, mean (SD)	885620	1.7 (1.8)
SOFA.SOFA Gesamt, mean (SD)	885612	9.1 (3.7)
SPAP, mean (SD)	790953	41.8 (12.7)
SpO2, mean (SD)	216505	95.1 (4.3)
SVRI, mean (SD)	893162	1488.3 (640.7)
SzvO2, mean (SD)	890497	79.4 (14.9)
Thrombozyten, mean (SD)	881825	217.3 (132.3)
Troponin, mean (SD)	890221	0.2 (0.5)
ZVD, mean (SD)	562353	11.4 (8.4)

List of parameters present for the NONCOVID cohort

	Missing	Overall
n		6343647
24h-Bilanz (Fluessigkeiten-Einfuhr vs -Ausfuhr), mean (SD)	6284785	52.0 (1512.6)
AF, mean (SD)	1512446	19.3 (6.4)
AF spontan, mean (SD)	3964921	10.3 (10.9)
Albumin, mean (SD)	6317427	490.4 (2209.2)
Amylase, mean (SD)	6340987	123.7 (593.3)
BE arteriell, mean (SD)	5862561	2.3 (4.3)

		Missing	Overall
Bicarbonat arteriell, mean (SD)		5861173	26.8 (4.4)
Bilirubin ges., mean (SD)		6284649	25.8 (57.2)
BMI, n (%)	1	0	61 (20.5)
	2		35 (11.8)
	3		34 (11.4)
	L		3 (1.0)
	M		38 (12.8)
	P		125 (42.1)
	nan		1 (0.3)
BNP, mean (SD)		6340220	7606.2 (13561.6)
CK, mean (SD)		6259645	250.0 (1214.2)
CK-MB, mean (SD)		6331213	54.3 (99.2)
clusterAlter, n (%)	70-79	0	59 (19.9)
	80-130		16 (5.4)
	<70		222 (74.7)
clusterGeschlecht, n (%)	M	0	205 (69.0)
	W		92 (31.0)
clusterKoerpergewicht, n (%)	65-75	0	42 (14.1)
	76-250		235 (79.1)
	<65		20 (6.7)
clusterKoerpergroesse, n (%)	180-185	0	85 (28.6)
	<180		194 (65.3)
	185		18 (6.1)
Compliance, mean (SD)		5033761	58.9 (40.0)
CRP, mean (SD)		6305612	1064.9 (778.8)
D-Dimere, mean (SD)		6336000	18133.2 (45764.6)

	Missing	Overall
DAP, mean (SD)	1424725	59.6 (13.2)
deltaP, mean (SD)	4823495	13.4 (4.7)
DPAP, mean (SD)	6209505	22.2 (8.0)
ECMO, mean (SD)	839541	0.0 (0.0)
etCO2, mean (SD)	6308698	41.0 (10.9)
Extrakorporaler Blutfluss, mean (SD)	6340555	4.0 (1.0)
Extrakorporaler Gasfluss (O2), mean (SD)	6337088	6.1 (10.0)
FiO2, mean (SD)	3996909	41.5 (14.8)
Gaszusammensetzung (%O2), mean (SD)	6336095	82.0 (24.7)
GOT, mean (SD)	6283863	146.0 (681.9)
GPT, mean (SD)	6284929	95.0 (269.8)
Haematokrit, mean (SD)	5795246	28.7 (4.9)
Haemoglobin, mean (SD)	5855449	5.8 (1.0)
Harnstoff, mean (SD)	6279895	10.6 (7.0)
HF, mean (SD)	979436	86.0 (17.9)
Horowitz-Quotient (ohne Temp-Korrektur), mean (SD)	5762619	319.4 (208.8)
I:E, mean (SD)	5154890	2.0 (1.3)
IL-6, mean (SD)	6336343	2246.9 (32357.1)
individuelles Tidalvolumen pro kg idealem Koerpergewicht, mean (SD)	6083961	401.7 (63.2)
Inhalatives NO, mean (SD)	6313089	16.0 (9.0)
INR, mean (SD)	6262724	1.3 (0.5)
iSOFA.HKL, mean (SD)	5174195	1.5 (1.5)
iSOFA.iSOFA Gesamt, mean (SD)	5174195	5.4 (4.2)
iSOFA.Leber, mean (SD)	5174195	0.4 (1.0)
iSOFA.Lunge, mean (SD)	5174195	1.3 (1.2)
iSOFA.Niere, mean (SD)	5174195	1.1 (1.6)
iSOFA.Thrombo, mean (SD)	5174195	0.5 (0.9)

		Missing	Overall
iSOFA.ZNS, mean (SD)		5174195	0.6 (1.3)
Koerperkerntemperatur, mean (SD)		1905052	36.7 (2.6)
Kreatinin, mean (SD)		6280568	102.4 (86.8)
Lagerungstherapie, mean (SD)		839541	0.0 (0.0)
Laktat arteriell, mean (SD)		5860610	1.4 (1.7)
LDH, mean (SD)		6312251	446.6 (803.4)
Leukozyten, mean (SD)		6273667	12.2 (10.1)
Lipase, mean (SD)		6324774	81.4 (283.6)
Lymphozyten absolut, mean (SD)		6341343	1.2 (0.7)
Lymphozyten prozentual, mean (SD)		6341340	0.1 (0.1)
MAP, mean (SD)		1389117	80.1 (16.0)
MPAP, mean (SD)		6203580	28.7 (9.7)
P EI, mean (SD)		4076392	19.7 (7.7)
paCO2 (ohne Temp-Korrektur), mean (SD)		6030250	41.9 (9.2)
paO2 (ohne Temp-Korrektur), mean (SD)		5992333	90.7 (33.6)
PCT, mean (SD)		6293951	2.8 (9.5)
PCWP, mean (SD)		6341358	17.4 (6.1)
PEEP, mean (SD)		4308983	7.0 (2.6)
PEEP eingestellt, mean (SD)		6084672	7.8 (2.1)
Phase, n (%)	0	0	35 (11.8)
	1		36 (12.1)
	2		226 (76.1)
pTT, mean (SD)		6260378	38.3 (15.2)
SaO2, mean (SD)		5862053	92.7 (11.2)
SAP, mean (SD)		1422645	124.3 (24.7)
SOFA.Punkte Blut, mean (SD)		6287648	0.5 (0.9)
SOFA.Punkte Leber, mean (SD)		6292186	0.4 (1.0)
SOFA.Punkte Lunge, mean (SD)		6288951	1.9 (0.9)
SOFA.Punkte Lunge (cal.), mean (SD)		6330760	1.7 (0.9)
SOFA.Punkte Niere, mean (SD)		6285317	3.6 (0.8)

	Missing	Overall
SOFA.Punkte ZNS, mean (SD)	6285532	0.7 (1.4)
SOFA.SOFA Gesamt, mean (SD)	6285317	7.7 (3.1)
SPAP, mean (SD)	6209424	40.2 (13.4)
SpO2, mean (SD)	1623454	96.5 (3.6)
SVRI, mean (SD)	6342431	1493.3 (703.2)
SzvO2, mean (SD)	6312342	72.4 (16.2)
Thrombozyten, mean (SD)	6273883	232.1 (146.4)
Troponin, mean (SD)	6337555	1.0 (3.5)
ZVD, mean (SD)	5143506	11.7 (8.9)

Encounters recorded for the ASIC study per Hospital

Hospital	Control Phase	Roll-In Phase	Quality Assurance Phase	Total
Hospital A	1081	358	2237	3676
Hospital B	929	175	467	1571
Hospital C	667	81	154	902
Hospital D	806	183	368	1360
Hospital E	350	61	75	486
Hospital F	584	84	10	678
Hospital G	2184	33	0	2217
Hospital F	5164	543	545	5164

List of Parameters explored during Hyperparameter Tuning

Parameter	Range
lambda:	1e-5 to 1e-2
alpha:	1e-5 to 1e-2
max depth:	2 to 6

Parameter	Range
gamma:	0 to 1.0
min_child_weight:	1 to 10
max_delta_step:	0 to 10
subsample:	0.6 to 1.0
colsample_bytree:	0.6 to 1.0
grow_policy:	depthwise or lossguide
scale_pos_weight:	1, 10, 30, 50, 100, 1000

Best Performing NON-COVID Model Parameters

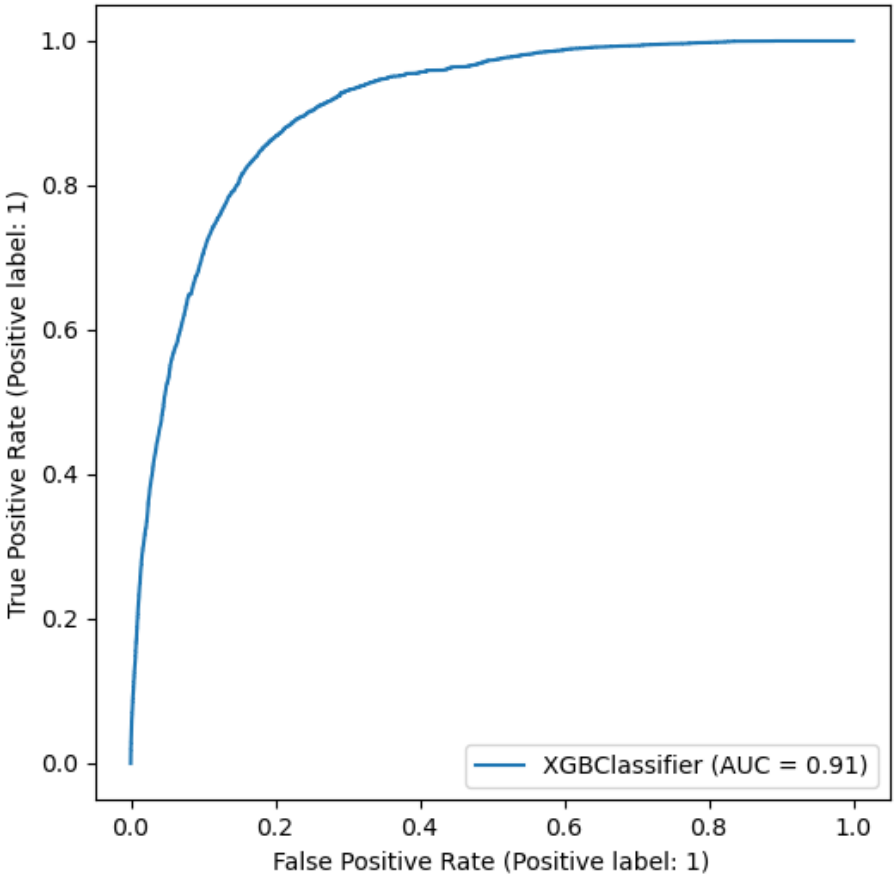


Figure 5.1: ROC Curve for the best performing NON-COVID model.

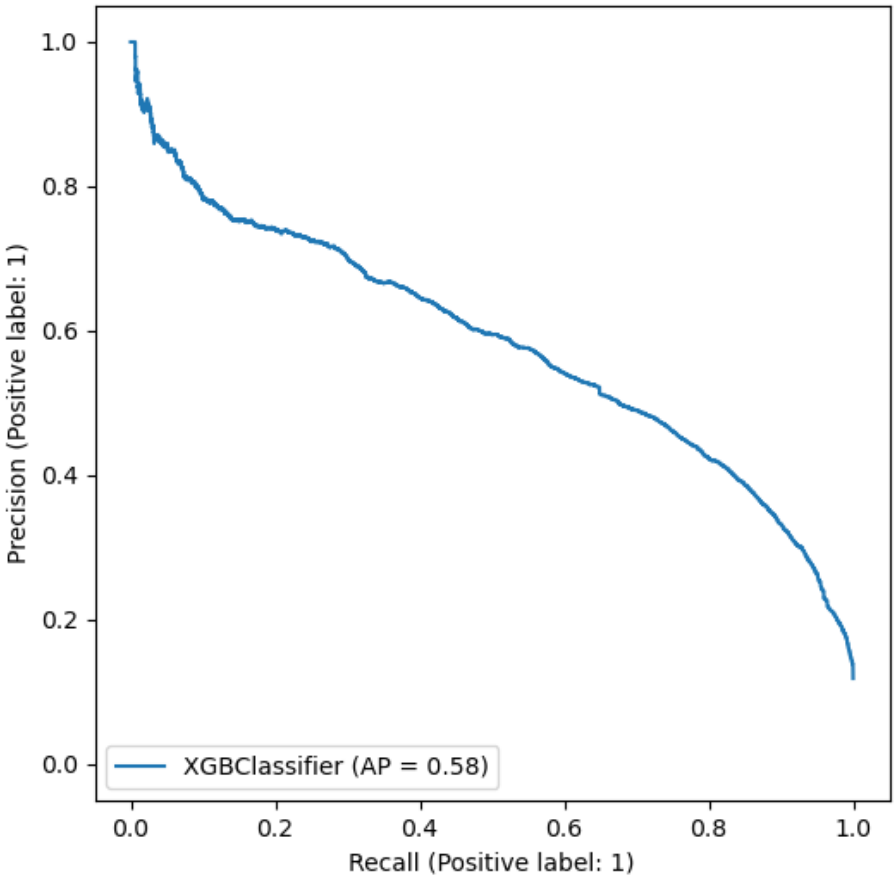


Figure 5.2: PR Curve for the best performing NON-COVID model.

Parameter	Value
alpha	0.009610670686699454
colsample_bytree	0.9063721890058604
gamma	0.9996996285423198
grow_policy	depthwise
lambda	0.008468086470697647
max_delta_step	6
max_depth	6
min_child_weight	8
objective	binary:logistic

Parameter	Value
scale_pos_weight	1
subsample	0.6521717498743645
tree_method	hist

Best Performing COVID Model Parameters

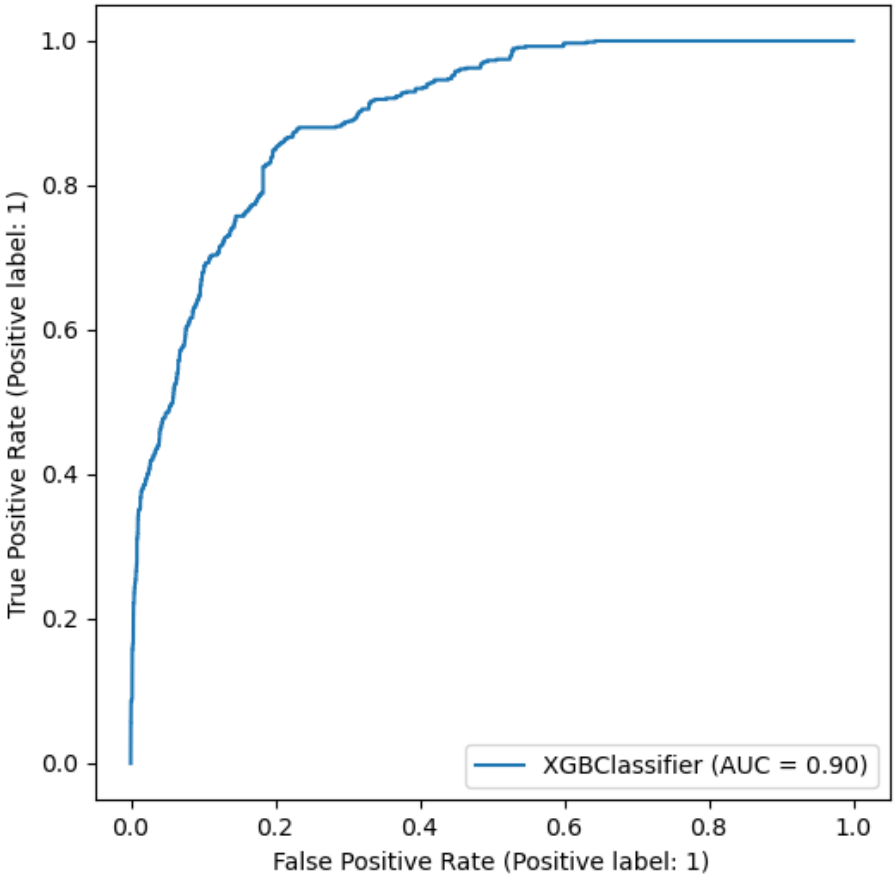


Figure 5.3: ROC Curve for the best performing COVID model.

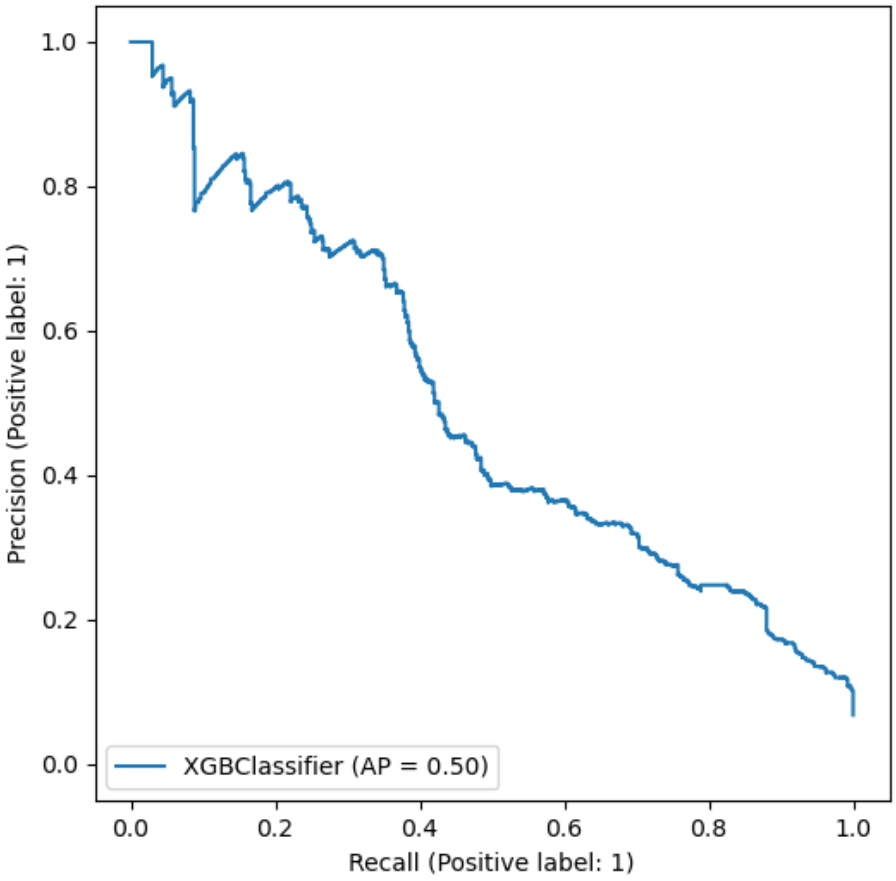


Figure 5.4: PR Curve for the best performing COVID model.

Parameter	Value
alpha	0.005599158832839225
colsample_bytree	0.694225541132862
gamma	0.4185214209985906
grow_policy	depthwise
lambda	0.009197310782529994
max_delta_step	4
max_depth	6
min_child_weight	9
n_boost_rounds	600

Parameter	Value
objective	binary:logistic
scale_pos_weight	1
subsample	0.6226869122104126
tree_method	hist

ROC and PR Curves for the Transfer Model

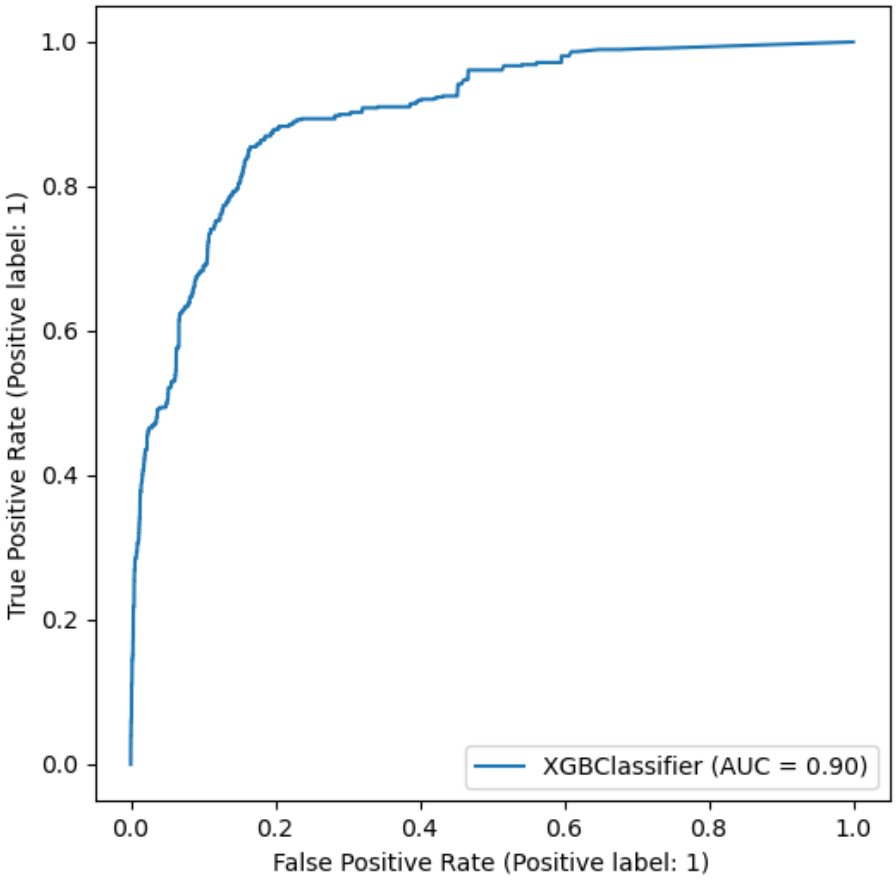


Figure 5.5: ROC Curve for the best performing transfer model.

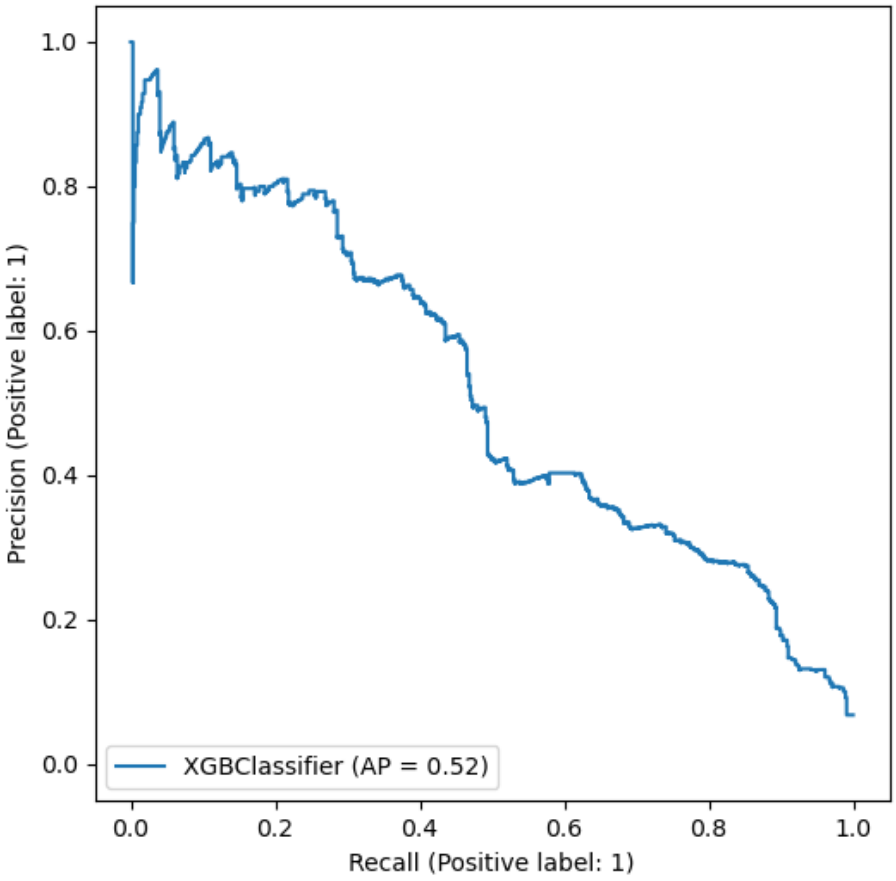


Figure 5.6: PR Curve for the best performing transfer model.

References

[1] Fan E, Brodie D, Slutsky AS. Acute Respiratory Distress Syndrome: Advances in Diagnosis and Treatment. JAMA 2018;319:698–710. https://doi.org/10.1001/jama.2017.21907.

[2] Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. JAMA 2016;315:788–800. https://doi.org/10.1001/jama.2016.0291.

[3] PA John E. Brodie MS, PA Ronald Johnson CCP, Lich BV. The Manual of Clinical Perfusion. Second edition. Perfusion.com; 2004.

[4] Zahar JR, Azoulay E, Klement E, De Lassence A, Lucet JC, Regnier B, et al. Delayed treatment contributes to mortality in ICU patients with severe active pulmonary tuberculosis and acute respiratory failure. Intensive Care Med 2001;27:513–20. https://doi.org/10.1007/s001340000849.

[5] Thille AW, Esteban A, Fernández-Segoviano P, Rodriguez J-M, Aramburu J-A, Peñuelas O, et al. Comparison of the Berlin Definition for Acute Respiratory Distress Syndrome with Autopsy. Am J Respir Crit Care Med 2013;187:761–7. https://doi.org/10.1164/rccm.201211-1981OC.

[6] Thille AW, Esteban A, Fernández-Segoviano P, Rodriguez J-M, Aramburu J-A, Vargas-Errázuriz P, et al. Chronology of histological lesions in acute respiratory distress syndrome with diffuse alveolar damage: A prospective cohort study of clinical autopsies. The Lancet Respiratory Medicine 2013;1:395–401. https://doi.org/10.1016/S2213-2600(13)70053-5.

[7] Marx G, Bickenbach J, Fritsch SJ, Kunze JB, Maassen O, Deffge S, et al. Algorithmic surveillance of ICU patients with acute respiratory distress syndrome (ASIC): Protocol for a multicentre stepped-wedge cluster randomised quality improvement strategy. BMJ Open 2021;11:e045589. https://doi.org/10.1136/bmjopen-2020-045589.

[8] Taoum A, Mourad-Chehade F, Amoud H. Early-warning of ARDS using novelty detection and data fusion. Comput Biol Med 2018;102:191–9. https://doi.org/10.1016/j.compbiomed.2018.09.030.

[9] Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care 2020;60:96–102. https://doi.org/10.1016/j.jcrc.2020.07.019.

[10] Singhal L, Garg Y, Yang P, Tabaie A, Wong AI, Mohammed A, et al. eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset Acute Respiratory Distress Syndrome (ARDS) among critically ill adults with COVID-19. PLOS ONE 2021;16:e0257056. https://doi.org/10.1371/journal.pone.0257056.

[11] Polzin R, Fritsch S, Sharafutdinov K, Bickenbach J, Marx G, Schuppert A. Predicting a sudden decrease in oxygenation in mechanically ventilated intensive care patients as a surrogate marker for acute respiratory distress syndrome n.d.

[12] Polzin R, Fritsch S, Sharafutdinov K, Marx G, Schuppert A. Diagnostic Expert Advisor: A platform for developing machine learning models on medical time-series data. SoftwareX 2023;23. https://doi.org/10.1016/j.softx.2023.101517.

[13] Sharafutdinov K. Virtual patient modeling for heterogeneous intensive care unit data for the support of artificial intelligence. Doctoral dissertation [RWTH Aachen University] 2023. https://doi.org/10.18154/RWTH-2023-05200.

[14] Sharafutdinov K, Bhat J, Fritsch S, Nikulina K, Samadi M, Polzin R, et al. Application of convex hull analysis for the evaluation of data heterogeneity between patient populations of different origin and implications of hospital bias in downstream machine-learning-based data processing: A comparison of 4 critical-care patient datasets. Frontiers in Big Data 2022;5:603429. https://doi.org/10.3389/fdata.2022.603429.

[15] Sharafutdinov K, Fritsch S, Iravani M, Farhadi P, Saffaran S, Bates D, et al. Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets. IEEE Open Journal of Engineering in Medicine and Biology 2023;PP:1–11. https://doi.org/10.1109/OJEMB.2023.3243190.

[16] Meyer J, Fritsch S, Sharafutdinov K, Nikulina K, Polzin R, Marx G, et al. Transfer Learning Across Diseases Opens a Novel Route Towards Pandemic Preparedness [Preprint] 2023. https://doi.org/10.21203/rs.3.rs-3349295/v1.

[17] Linden T, Ku C, Wendland K, Sharafutdinov K, Polzin R, Schuppert A, et al. Survival Multi-Modal Neural Ordinary Differential Equations for Mortality Prediction of Patients with Severe Lung Disease [Unpublished Manuscript] n.d.

[18] Hagens LA, Van der Ven FLIM, Heijnen NFL, Smit MR, Gietema HA, Gerretsen SC, et al. Improvement of an interobserver agreement of ARDS diagnosis by adding additional imaging and a confidence scale. Front Med (Lausanne) 2022;9:950827. https://doi.org/10.3389/fmed.2022.950827.

[19] Smeden M van, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: Diagnosis versus prognosis. Journal of Clinical Epidemiology 2021;132:142–5. https://doi.org/10.1016/j.jclinepi.2021.01.009.

[20] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683–90. https://doi.org/10.1136/heartjnl-2011-301246.

[21] Holmboe ES, Durning SJ. Assessing clinical reasoning: Moving from in vitro to in vivo. Diagnosis (Berl) 2014;1:111–7. https://doi.org/10.1515/dx-2013-0029.

[22] Herasevich S, Pinevich Y, Lindroth HL, Herasevich V, Pickering BW, Barwise AK. Who needs clinician attention first? A qualitative study of critical care clinicians' needs that enable the prioritization of care for populations of acutely ill patients. International Journal of Medical Informatics 2023;177:105118. https://doi.org/10.1016/j.ijmedinf.2023.105118.

[23] Elektrofahrzeuge, Solaranlagen und saubere Energie n.d. https://www.tesla.com/ (accessed February 19, 2024).

[24] OpenAI. Sora n.d. http://web.archive.org/web/20240220000958/https://openai.com/sora (accessed February 19, 2024).

[25] AI will transform science — now researchers must tame it. Nature 2023;621:658–8. https://doi.org/10.1038/d41586-023-02988-6.

[26] Van Noorden R, Perkel JM. AI and science: What 1,600 researchers think. Nature 2023;621:672–5. https://doi.org/10.1038/d41586-023-02980-0.

[27] What is Artificial Intelligence (AI) ? | IBM n.d. http://web.archive.org/web/20240620110500/https://www.ibm.com/topics/artificial-intelligence (accessed June 22, 2023).

[28] Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: Current trends and future possibilities. Br J Gen Pract 2018;68:143–4. https://doi.org/10.3399/bjgp18X695213.

[29] Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. Annals of Surgery 2018;268:70. https://doi.org/10.1097/SLA.0000000000002693.

[30] Tzeng I-S, Hsieh P-C, Su W-L, Hsieh T-H, Chang S-C. Artificial Intelligence-Assisted Chest X-ray for the Diagnosis of COVID-19: A Systematic Review and Meta-Analysis. Diagnostics (Basel) 2023;13:584. https://doi.org/10.3390/diagnostics13040584.

[31] Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today 2021;26:80–93. https://doi.org/10.1016/j.drudis.2020.10.010.

[32] Teatini A, Pelanis E, Aghayan DL, Kumar RP, Palomar R, Fretland ÅA, et al. The effect of intraoperative imaging on surgical navigation for laparoscopic liver resection surgery. Sci Rep 2019;9:18687. https://doi.org/10.1038/s41598-019-54915-3.

[33] May M. The next generation of robotic surgery is emerging: But is it better than a human? Nature Medicine 2024;30:2–5. https://doi.org/10.1038/s41591-023-02740-7.

[34] Holohan M, Fiske A. "Like I'm Talking to a Real Person": Exploring the Meaning of Transference for the Use and Design of AI-Based Applications in Psychotherapy. Front Psychol 2021;12:720476. https://doi.org/10.3389/fpsyg.2021.720476.

[35] Pianykh OS, Guitron S, Parke D, Zhang C, Pandharipande P, Brink J, et al. Improving healthcare operations management with machine learning. Nat Mach Intell 2020;2:266–73. https://doi.org/10.1038/s42256-020-0176-3.

[36] Kause J, Smith G, Prytherch D, Parr M, Flabouris A, Hillman K, et al. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom–the ACADEMIA study. Resuscitation 2004;62:275–82. https://doi.org/10.1016/j.resuscitation.2004.05.016.

[37] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016;375:2293–7. https://doi.org/10.1056/NEJMsb1609216.

[38] US Food & Drug Administration. Webinar: Framework for FDA's Real-World Evidence Program – Mar 15, 2019 2019. http://web.archive.org/web/20241008010030/https://www.fda.gov/drugs/webinar-framework-fdas-real-world-evidence-program-mar-15-2019 (accessed February 24, 2024).

[39] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. Npj Digit Med 2020;3:1–10. https://doi.org/10.1038/s41746-020-0221-y.

[40] Omididan Z, Hadianfar A. The role of clinical decision support systems in healthcare (1980-2010): A systematic review study. Jentashapir Sceintific-Research Quarterly 2011;2:125–34.

[41] Shahmoradi L, Safdari R, Ahmadi H, Zahmatkeshan M. Clinical decision support systems-based interventions to improve medication outcomes: A systematic literature review on features and effects. Med J Islam Repub Iran 2021;35:27. https://doi.org/10.47176/mjiri.35.27.

[42] Dias D, Paulo Silva Cunha J. Wearable Health Devices—Vital Sign Monitoring, Systems and Technologies. Sensors 2018;18:2414. https://doi.org/10.3390/s18082414.

[43] Berner ES, editor. Clinical Decision Support Systems. New York, NY: Springer New York; 2007. https://doi.org/10.1007/978-0-387-38319-4.

[44] Kundu S. AI in medicine must be explainable. Nat Med 2021;27:1328–8. https://doi.org/10.1038/s41591-021-01461-z.

[45] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–53. https://doi.org/10.1126/science.aax2342.

[46] Kilsdonk E, Peute LW, Riezebos RJ, Kremer LC, Jaspers MWM. Uncovering healthcare practitioners' information processing using the think-aloud method: From paper-based guideline to clinical decision support system. Int J Med Inform 2016;86:10–9. https://doi.org/10.1016/j.ijmedinf.2015.11.011.

[47] Goddard K, Roudsari A, Wyatt JC. Automation bias - a hidden issue for clinical decision support system use. Stud Health Technol Inform 2011;164:17–22.

[48] Kabachinski J. A look at clinical decision support systems. Biomed Instrum Technol 2013;47:432–4. https://doi.org/10.2345/0899-8205-47.5.432.

[49] Drew BJ, Califf RM, Funk M, Kaufman ES, Krucoff MW, Laks MM, et al. Practice standards for electrocardiographic monitoring in hospital settings. Circulation 2004;110:2721–46. https://doi.org/10.1161/01.CIR.0000145144.56673.59.

[50] Chambrin MC. Alarms in the intensive care unit: How can the number of false alarms be reduced? Crit Care 2001;5:184–8. https://doi.org/10.1186/cc1021.

[51] Siebig S, Kuhls S, Imhoff M, Gather U, Schölmerich J, Wrede CE. Intensive care unit alarms—How many do we need?*. Critical Care Medicine 2010;38:451. https://doi.org/10.1097/CCM.0b013e3181cb0888.

[52] Drew BJ, Harris P, Zègre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, et al. Insights into the Problem of Alarm Fatigue with Physiologic Monitor Devices: A Comprehensive Observational Study of Consecutive Intensive Care Unit Patients. PLOS ONE 2014;9:e110274. https://doi.org/10.1371/journal.pone.0110274.

[53] Donchin Y, Seagull FJ. The hostile environment of the intensive care unit. Current Opinion in Critical Care 2002;8:316.

[54] Poncette A-S, Wunderlich MM, Spies C, Heeren P, Vorderwülbecke G, Salgado E, et al. Patient Monitoring Alarms in an Intensive Care Unit: Observational Study With Do-It-Yourself Instructions. J Med Internet Res 2021;23:e26494. https://doi.org/10.2196/26494.

[55] Sendelbach S, Funk M. Alarm fatigue: A patient safety concern. AACN Adv Crit Care 2013;24:378-386; quiz 387-388. https://doi.org/10.1097/NCI.0b013e3182a903f9.

[56] Cvach M. Monitor alarm fatigue: An integrative review. Biomed Instrum Technol 2012;46:268–77. https://doi.org/10.2345/0899-8205-46.4.268.

[57] Torabizadeh C, Yousefinya A, Zand F, Rakhshan M, Fararooei M. A nurses' alarm fatigue questionnaire: Development and psychometric properties. J Clin Monit Comput 2017;31:1305–12. https://doi.org/10.1007/s10877-016-9958-x.

[58] Turmell JW, Coke L, Catinella R, Hosford T, Majeski A. Alarm Fatigue: Use of an Evidence-Based Alarm Management Strategy. J Nurs Care Qual 2017;32:47–54. https://doi.org/10.1097/NCQ.0000000000000223.

[59] Oliveira AEC de, Machado AB, Santos EDD, Almeida ÉB de. Alarm fatigue and the implications for patient safety. Rev Bras Enferm 2018;71:3035–40. https://doi.org/10.1590/0034-7167-2017-0481.

[60] Keller JP. Clinical alarm hazards: A "top ten" health technology safety concern. Journal of Electrocardiology 2012;45:588–91. https://doi.org/10.1016/j.jelectrocard.2012.08.050.

[61] ECRI Institute Releases Top 10 Health Technology Hazards Report for 2014 2016. https://web.archive.org/web/20160305005415/https://www.ecri.org/press/Pages/2014-Top-10-Health-Technology-Hazards-Report.aspx (accessed February 23, 2024).

[62] Laennec RTH. A treatise on the diseases of the chest: In which they are described according to their anatomical characters, and their diagnosis established on a new principle by means of acoustic instruments. 1821.

[63] Murray JF, Matthay MA, Luce JM, Flick MR. An expanded definition of the adult respiratory distress syndrome. Am Rev Respir Dis 1988;138:720–3. https://doi.org/10.1164/ajrccm/138.3.720.

[64] Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, et al. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. Am J Respir Crit Care Med 1994;149:818–24. https://doi.org/10.1164/ajrccm.149.3.7509706.

[65] ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: The Berlin Definition. JAMA 2012;307:2526–33. https://doi.org/10.1001/jama.2012.5669.

[66] Horovitz JH, Carrico CJ, Shires GT. Pulmonary Response to Major Injury. Archives of Surgery 1974;108:349–55. https://doi.org/10.1001/archsurg.1974.01350270079014.

[67] Feiner JR, Weiskopf RB. Evaluating Pulmonary Function: An Assessment of PaO2/FIO2. Crit Care Med 2017;45:e40–8. https://doi.org/10.1097/CCM.0000000000002017.

[68] Villar J, Pérez-Méndez L, Kacmarek RM. The Berlin definition met our needs: no. Intensive Care Med 2016;42:648–50. https://doi.org/10.1007/s00134-016-4242-6.

[69] Zbiral M, Weber M, König S, Kraft F, Ullrich R, Krenn K. Usefulness and limitations of the acute respiratory distress syndrome definitions in non-intubated patients. A narrative review. Front Med (Lausanne) 2023;10:1088709. https://doi.org/10.3389/fmed.2023.1088709.

[70] Fröhlich S, Murphy N, Doolan A, Ryan O, Boylan J. Acute respiratory distress syndrome: Underrecognition by clinicians. Journal of Critical Care 2013;28:663–8. https://doi.org/10.1016/j.jcrc.2013.05.012.

[71] Summers C, Singh NR, Worpole L, Simmonds R, Babar J, Condliffe AM, et al. Incidence and recognition of acute respiratory distress syndrome in a UK intensive care unit. Thorax 2016;71:1050–1. https://doi.org/10.1136/thoraxjnl-2016-208402.

[72] Confalonieri M, Salton F, Fabiano F. Acute respiratory distress syndrome. European Respiratory Review 2017;26. https://doi.org/10.1183/16000617.0116-2016.

[73] Brun-Buisson C, Minelli C, Bertolini G, Brazzi L, Pimentel J, Lewandowski K, et al. Epidemiology and outcome of acute lung injury in European intensive care units. Intensive Care Med 2004;30:51–61. https://doi.org/10.1007/s00134-003-2022-6.

[74] Caser EB, Zandonade E, Pereira E, Gama AMC, Barbas CSV. Impact of Distinct Definitions of Acute Lung Injury on Its Incidence and Outcomes in Brazilian ICUs: Prospective Evaluation of 7,133 Patients*. Critical Care Medicine 2014;42:574. https://doi.org/10.1097/01.ccm.0000435676.68435.56.

[75] Erickson SE, Martin GS, Davis JL, Matthay MA, Eisner MD, Network for the NNA. Recent trends in acute lung injury mortality: 1996–2005*. Critical Care Medicine 2009;37:1574. https://doi.org/10.1097/CCM.0b013e31819fefdf.

[76] Spieth PM, Güldner A, Gama de Abreu M. Akutes Lungenversagen. Anaesthesist 2017;66:539–52. https://doi.org/10.1007/s00101-017-0337-x.

[77] Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, et al. Incidence and outcomes of acute lung injury. N Engl J Med 2005;353:1685–93. https://doi.org/10.1056/NEJMoa050333.

[78] Low blood oxygen (hypoxemia) n.d. https://www.mayoclinic.org/symptoms/hypoxemia/basics/definition/sym-20050930 (accessed March 8, 2024).

[79] Qadir N, Chen J-T. Adjunctive Therapies in ARDS: The Disconnect Between Clinical Trials and Clinical Practice. CHEST 2020;157:1405–6. https://doi.org/10.1016/j.chest.2020.03.022.

[80] Diamond M, Peniston HL, Sanghavi DK, Mahapatra S. Acute Respiratory Distress Syndrome. StatPearls, Treasure Island (FL): StatPearls Publishing; 2024.

[81] Bellani G, Pham T, Laffey JG. Missed or delayed diagnosis of ARDS: A common and serious problem. Intensive Care Med 2020;46:1180–3. https://doi.org/10.1007/s00134-020-06035-0.

[82] NVIDIA GeForce RTX 4090 Specs 2024. https://www.techpowerup.com/gpu-specs/geforce-rtx-4090.c3889 (accessed February 18, 2024).

[83] Home - | TOP500 n.d. https://www.top500.org/ (accessed February 18, 2024).

[84] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners 2020. http://arxiv.org/abs/2005.14165 (accessed March 14, 2024).

[85] Microsoft announces new supercomputer, lays out vision for future AI work n.d. https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/ (accessed March 14, 2024).

[86] Kurz M, Offenhäuser P, Viola D, Shcherbakov O, Resch M, Beck A. Deep reinforcement learning for computational fluid dynamics on HPC systems. Journal of Computational Science 2022;65:101884. https://doi.org/10.1016/j.jocs.2022.101884.

[87] Lusk MT, Mattsson AE. High-performance computing for materials design to advance energy science. MRS Bulletin 2011;36:169–74. https://doi.org/10.1557/mrs.2011.30.

[88] Pyzer-Knapp EO, Pitera JW, Staar PWJ, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. Npj Comput Mater 2022;8:1–9. https://doi.org/10.1038/s41524-022-00765-z.

[89] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.

[90] Kim M, Yun J, Cho Y, Shin K, Jang R, Bae H, et al. Deep Learning in Medical Imaging. Neurospine 2019;16:657–68. https://doi.org/10.14245/ns.1938396.198.

[91] Niehoff JH, Kalaitzidis J, Kroeger JR, Schoenbeck D, Borggrefe J, Michael AE. Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays. Sci Rep 2023;13:3680. https://doi.org/10.1038/s41598-023-30521-2.

[92] Shuhaiber JH. Augmented Reality in Surgery. Archives of Surgery 2004;139:170–4. https://doi.org/10.1001/archsurg.139.2.170.

[93] Nashwan AJ, Abujaber AA, Choudry H. Embracing the future of physician-patient communication: GPT-4 in gastroenterology. Gastroenterology & Endoscopy 2023;1:132–5. https://doi.org/10.1016/j.gande.2023.07.004.

[94] Lamichhane B. Evaluation of ChatGPT for NLP-based Mental Health Applications 2023. https://doi.org/10.48550/arXiv.2303.15727.

[95] Koch M, Arlandini C, Antonopoulos G, Baretta A, Beaujean P, Bex GJ, et al. HPC+ in the medical field: Overview and current examples. Technology and Health Care 2023;31:1509–23. https://doi.org/10.3233/THC-229015.

[96] Pirracchio R. Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project. In: MIT Critical Data, editor. Secondary Analysis of Electronic Health Records, Cham (CH): Springer; 2016.

[97] Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: A mortality prediction case study. Proceedings of the 2nd Machine Learning for Healthcare Conference, PMLR; 2017, p. 361–76.

[98] Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. Journal of Biomedical Informatics 2018;83:112–34. https://doi.org/10.1016/j.jbi.2018.04.007.

[99] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. Npj Digital Med 2018;1:1–10. https://doi.org/10.1038/s41746-018-0029-1.

[100] Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data 2019;6:96. https://doi.org/10.1038/s41597-019-0103-9.

[101] Hassler AP, Menasalvas E, García-García FJ, Rodríguez-Mañas L, Holzinger A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Mak 2019;19:33. https://doi.org/10.1186/s12911-019-0747-6.

[102] Lin J-H, Haug PJ. Data Preparation Framework for Preprocessing Clinical Data in Data Mining. AMIA Annu Symp Proc 2006;2006:489–93.

[103] Bertsimas D, Orfanoudaki A, Pawlowski C. Imputation of clinical covariates in time series. Mach Learn 2021;110:185–248. https://doi.org/10.1007/s10994-020-05923-2.

[104] Luo Y. Evaluating the state of the art in missing data imputation for clinical data. Briefings in Bioinformatics 2022;23:bbab489. https://doi.org/10.1093/bib/bbab489.

[105] Jarrett D, Yoon J, Bica I, Qian Z, Ercole A, van der Schaar M. Clairvoyance: A Pipeline Toolkit for Medical Time Series 2023. http://arxiv.org/abs/2310.18688 (accessed February 25, 2024).

[106] Naul B, van der Walt S, Crellin-Quick A, Bloom JS, Pérez F. Cesium: Open-Source Platform for Time-Series Inference 2016. https://doi.org/10.48550/arXiv.1609.04504.

[107] Tavenard R, Faouzi J, Vandewiele G, Divo F, Androz G, Holtz C, et al. Tslearn, A Machine Learning Toolkit for Time Series Data. Journal of Machine Learning Research 2020;21:1–6.

[108] Burns DM, Whyne CM. Seglearn: A Python Package for Learning Sequences and Time Series. Journal of Machine Learning Research 2018;19:1–7.

[109] Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 2018;307:72–7. https://doi.org/10.1016/j.neucom.2018.03.067.

[110] Löning M, Bagnall A, Ganesh S, Kazakov V, Lines J, Király FJ. Sktime: A Unified Interface for Machine Learning with Time Series 2019. https://doi.org/10.48550/arXiv.1909.07872.

[111] Faouzi J, Janati H. Pyts: A Python Package for Time Series Classification. Journal of Machine Learning Research 2020;21:1–6.

[112] Knowles R, Mateen BA, Yehudi Y. We need to talk about the lack of investment in digital research infrastructure. Nat Comput Sci 2021;1:169–71. https://doi.org/10.1038/s43588-021-00048-5.

[113] Zhou S, Brunke L, Tao A, Hall AW, Bejarano FP, Panerati J, et al. What is the Impact of Releasing Code with Publications? Statistics from the Machine Learning, Robotics, and Control Communities 2023. http://arxiv.org/abs/2308.10008 (accessed February 25, 2024).

[114] Vandewalle P. Code availability for image processing papers: A status update - KU Leuven 2019. https://lirias.kuleuven.be/retrieve/541895 (accessed February 16, 2024).

[115] Lee BD. Ten simple rules for documenting scientific software. PLOS Computational Biology 2018;14:e1006561. https://doi.org/10.1371/journal.pcbi.1006561.

[116] Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology 2013;9:e1003285. https://doi.org/10.1371/journal.pcbi.1003285.

[117] Archenaa J, Anita EAM. A Survey of Big Data Analytics in Healthcare and Government. Procedia Computer Science 2015;50:408–13. https://doi.org/10.1016/j.procs.2015.04.021.

[118] Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med 2020;8:2050312120934839. https://doi.org/10.1177/2050312120934839.

[119] Goodyear MDE, Krleza-Jeric K, Lemmens T. The Declaration of Helsinki. BMJ 2007;335:624–5. https://doi.org/10.1136/bmj.39339.610000.BE.

[120] Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. BMJ 2015;350:h391. https://doi.org/10.1136/bmj.h391.

[121] Coronavirus: Bayerische Behörden bestätigen ersten Fall in Deutschland. Der Spiegel 2020.

[122] Hussain M, Syed SK, Fatima M, Shaukat S, Saadullah M, Alqahtani AM, et al. Acute Respiratory Distress Syndrome and COVID-19: A Literature Review. Journal of Inflammation Research 2021;14:7225. https://doi.org/10.2147/JIR.S334043.

[123] Recital 26 - Not applicable to anonymous data 2018. https://gdpr.eu/recital-26-not-applicable-to-anonymous-data/ (accessed February 14, 2024).

[124] Art. 4 GDPR - Definitions 2018. https://gdpr.eu/article-4-definitions/ (accessed February 14, 2024).

[125] Eke D, Aasebø IEJ, Akintoye S, Knight W, Karakasidis A, Mikulan E, et al. Pseudonymisation of neuroimages and data protection: *Increasing Access to Data While Retaining Scientific Utility*. Neuroimage: Reports 2021;1:100053. https://doi.org/10.1016/j.ynirp.2021.100053.

[126] Samarati P, Sweeney L. Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression, 1998.

[127] El Emam K, Dankar FK, Issa R, Jonker E, Amyot D, Cogo E, et al. A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. Journal of the American Medical Informatics Association 2009;16:670–82. https://doi.org/10.1197/jamia.M3144.

[128] Olatunji IE, Rauch J, Katzensteiner M, Khosla M. A Review of Anonymization for Healthcare Data. Big Data 2022. https://doi.org/10.1089/big.2021.0169.

[129] Gokhale M, Stürmer T, Buse JB. Real-world evidence: The devil is in the detail. Diabetologia 2020;63:1694–705. https://doi.org/10.1007/s00125-020-05217-1.

[130] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.

[131] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x.

[132] Evans TW, Nava S, Mata GV, Guidet B, Estenssoro E, Fowler R, et al. Critical care rationing: International comparisons. Chest 2011;140:1618–24. https://doi.org/10.1378/chest.11-0957.

[133] Mai H, Mai H. Kartenbetrug in Deutschland: Geringer Anteil, aber hohe Kosten 2018. https://www.dbresearch.de/PROD/RPS_DE-PROD/PROD0000000000484136.pdf.

[134] Li D-C, Liu C-W, Hu SC. A learning method for the class imbalance problem with medical data sets. Computers in Biology and Medicine 2010;40:509–18. https://doi.org/10.1016/j.compbiomed.2010.03.005.

[135] Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial Intelligence in Medicine 2019;101:101723. https://doi.org/10.1016/j.artmed.2019.101723.

[136] Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. Int Conf Affect Comput Intell Interact Workshops 2013;2013:245–51. https://doi.org/10.1109/ACII.2013.47.

[137] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432.

[138] Huang T-H, Fan B, Rothschild MF, Hu Z-L, Li K, Zhao S-H. MiRFinder: An improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics 2007;8:341. https://doi.org/10.1186/1471-2105-8-341.

[139] Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. MONAI: An open-source framework for deep learning in healthcare 2022. http://arxiv.org/abs/2211.02701 (accessed May 3, 2023).

[140] Yang C, Wu Z, Jiang P, Lin Z, Gao J, Danek BP, et al. PyHealth: A Deep Learning Toolkit for Healthcare Applications. KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM Association for Computing Machinery; 2023. https://doi.org/10.1145/3580305.3599178.

[141] Anzt H, Bach F, Druskat S, Löffler F, Loewe A, Renard B, et al. An environment for sustainable research software in Germany and beyond: Current state, open challenges, and call for action. F1000Research 2021;9. https://doi.org/10.12688/f1000research.23224.2.

[142] Zaharia M, Chen A, Davidson A, Ghodsi A, Hong SA, Konwinski A, et al. Accelerating the Machine Learning Lifecycle with MLflow. IEEE Data Eng Bull 2018;41:39–45.

[143] Neptune.ai. Neptune.ai : Experiment Tracking and Model Registry 2022. https://neptune.ai.

[144] Biewald L. Experiment Tracking with Weights and Biases 2020. https://www.wandb.com/.

[145] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems 2015.

[146] Horsfall D, Cool J, Hettrick S, Pisco AO, Hong NC, Haniffa M. Research software engineering accelerates the translation of biomedical research for health. Nat Med 2023;29:1313–6. https://doi.org/10.1038/s41591-023-02353-0.

[147] Flask 2023. https://github.com/pallets/flask (accessed April 27, 2023).

[148] Pandas-dev/pandas: Pandas 2020. https://doi.org/10.5281/zenodo.3509134.

[149] Yoo AB, Jette MA, Grondona M. SLURM: Simple Linux Utility for Resource Management. In: Feitelson D, Rudolph L, Schwiegelshohn U, editors. Job Scheduling Strategies for Parallel Processing, Berlin, Heidelberg: Springer; 2003, p. 44–60. https://doi.org/10.1007/10968987_3.

[150] Joblib Development Team. Joblib: Running Python functions as pipeline jobs 2020. https://joblib.readthedocs.io/.

[151] Bokeh Development Team. Bokeh: Python library for interactive visualization 2018. https://bokeh.pydata.org/en/latest/.

[152] PyGWalker 2023. https://github.com/Kanaries/pygwalker (accessed April 27, 2023).

[153] Tableau: Business Intelligence and Analytics Software n.d. https://www.tableau.com/node/62770 (accessed April 28, 2023).

[154] Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

[155] Microsoft Corporation. Microsoft Excel 2018.

[156] Toolbox SM et al. Matlab. Mathworks Inc 1993.

[157] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.

[158] Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas, IOS Press; 2016, p. 87–90.

[159] pmc. Apache Arrow 9.0.0 Release 2022. https://arrow.apache.org/blog/2022/08/16/9.0.0-release/ (accessed April 28, 2023).

[160] Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 2007;9:90–5. https://doi.org/10.1109/MCSE.2007.55.

[161] Fritsch S, Maassen O, Riedel M. Artificial Intelligence: Infrastructures and Prerequisites at European Level. Anasthesiol Intensivmed Notfallmed Schmerzther 2022;57:172–84. https://doi.org/10.1055/a-1423-8052.

[162] Maassen O, Fritsch S, Palm J, Deffge S, Kunze J, Marx G, et al. Future Medical Artificial Intelligence Application Requirements and Expectations of Physicians in German University Hospitals: Web-Based Survey. J Med Internet Res 2021;23:e26646. https://doi.org/10.2196/26646.

[163] Matthay MA, Zemans RL, Zimmerman GA, Arabi YM, Beitler JR, Mercat A, et al. Acute respiratory distress syndrome. Nat Rev Dis Primers 2019;5:1–22. https://doi.org/10.1038/s41572-019-0069-0.

[164] Kor DJ, Carter RE, Park PK, Festic E, Banner-Goodspeed VM, Hinds R, et al. Effect of Aspirin on Development of ARDS in At-Risk Patients Presenting to the Emergency Department: The LIPS-A Randomized Clinical Trial. JAMA 2016;315:2406–14. https://doi.org/10.1001/jama.2016.6330.

[165] Gajic O, Dabbagh O, Park PK, Adesanya A, Chang SY, Hou P, et al. Early Identification of Patients at Risk of Acute Lung Injury. Am J Respir Crit Care Med 2011;183:462–70. https://doi.org/10.1164/rccm.201004-0549OC.

[166] Trillo-Alvarez C, Cartin-Ceba R, Kor DJ, Kojicic M, Kashyap R, Thakur S, et al. Acute lung injury prediction score: Derivation and validation in a population-based sample. Eur Respir J 2011;37:604–9. https://doi.org/10.1183/09031936.00036810.

[167] Soto GJ, Kor DJ, Park PK, Hou PC, Kaufman DA, Kim M, et al. Lung Injury Prediction Score in Hospitalized Patients at risk of Acute Respiratory Distress Syndrome. Crit Care Med 2016;44:2182–91. https://doi.org/10.1097/CCM.0000000000002001.

[168] Brown LM, Calfee CS, Matthay MA, Brower RG, Thompson BT, Checkley W. A simple classification model for hospital mortality in patients with acute lung injury managed with lung protective ventilation. Crit Care Med 2011;39:2645–51. https://doi.org/10.1097/CCM.0b013e3182266779.

[169] Villar J, González-Martín JM, Hernández-González J, Armengol MA, Fernández C, Martín-Rodríguez C, et al. Predicting ICU mortality in acute respiratory distress syndrome patients using machine learning: The Predicting Outcome and STratifiCation of severity in ARDS (POSTCARDS) Study. Critical Care Medicine 2023;51:1638–49. https://doi.org/10.1097/CCM.0000000000006030.

[170] Wang Z, Zhang L, Huang T, Yang R, Cheng H, Wang H, et al. Developing an explainable machine learning model to predict the mechanical ventilation duration of patients with ARDS in intensive care units. Heart & Lung 2023;58:74–81. https://doi.org/10.1016/j.hrtlng.2022.11.005.

[171] Sayed M, Riaño D, Villar J. Novel criteria to classify ARDS severity using a machine learning approach. Crit Care 2021;25:150. https://doi.org/10.1186/s13054-021-03566-w.

[172] Afshar M, Joyce C, Oakey A, Formanek P, Yang P, Churpek MM, et al. A Computable Phenotype for Acute Respiratory Distress Syndrome Using Natural Language Processing and Machine Learning. AMIA Annu Symp Proc 2018;2018:157–65.

[173] Zaglam N, Jouvet P, Flechelles O, Emeriaud G, Cheriet F. Computer-aided diagnosis system for the Acute Respiratory Distress Syndrome from chest radiographs. Computers in Biology and Medicine 2014;52:41–8. https://doi.org/10.1016/j.compbiomed.2014.06.006.

[174] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space 2013. http://arxiv.org/abs/1301.3781 (accessed March 1, 2024).

[175] Serpa Neto A, Deliberato RO, Johnson AEW, Bos LD, Amorim P, Pereira SM, et al. Mechanical power of ventilation is associated with mortality in critically ill patients: An analysis of patients in two observational cohorts. Intensive Care Med 2018;44:1914–22. https://doi.org/10.1007/s00134-018-5375-6.

[176] DIMDI - ICD-10-GM Version 2020 n.d. https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2020/block-u00-u49.htm (accessed March 7, 2024).

[177] Leontjeva A, Kuzovkin I. Combining Static and Dynamic Features for Multivariate Sequence Classification. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016, p. 21–30. https://doi.org/10.1109/DSAA.2016.10.

[178] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.

[179] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

[180] Stone RAO Jerome Friedman. Classification and Regression Trees. New York: Routledge; 2017. https://doi.org/10.1201/9781315139470.

[181] Predicting Red Hat Business Value n.d. https://kaggle.com/competitions/predicting-red-hat-business-value (accessed January 26, 2024).

[182] OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction n.d. https://kaggle.com/competitions/stanford-covid-vaccine (accessed January 26, 2024).

[183] Flavours of Physics n.d. https://kaggle.com/competitions/flavours-of-physics (accessed January 26, 2024).

[184] Su Y, Yuan D, Chen DG, Ng RH, Wang K, Choi J, et al. Multiple early factors anticipate post-acute COVID-19 sequelae. Cell 2022;185:881–895.e20. https://doi.org/10.1016/j.cell.2022.01.014.

[185] Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2020;2:283–8. https://doi.org/10.1038/s42256-020-0180-7.

[186] Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: A population-based, diagnostic study. The Lancet Oncology 2020;21:222–32. https://doi.org/10.1016/S1470-2045(19)30738-7.

[187] Seni G, Elder JF. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Cham: Springer International Publishing; 2010. https://doi.org/10.1007/978-3-031-01899-2.

[188] Cawley GC, Talbot NLC. On Over-fittingg inModel Selectionn andSubsequent Selection Biass inPerformance Evaluation. Journal of Machine Learning Research 2010;11.

[189] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework 2019. https://doi.org/10.48550/arXiv.1907.10902.

[190] Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms 2012. https://doi.org/10.48550/arXiv.1206.2944.

[191] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

[192] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access 2021;9:109960–75. https://doi.org/10.1109/ACCESS.2021.3102399.

[193] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 2017;73:220–39. https://doi.org/10.1016/j.eswa.2016.12.035.

[194] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data 2019;6:27. https://doi.org/10.1186/s40537-019-0192-5.

[195] Batista G, Prati R, Monard M-C. A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. SIGKDD Explorations 2004;6:20–9. https://doi.org/10.1145/1007730.1007735.

[196] IPBES. Workshop Report on Biodiversity and Pandemics of the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES). Zenodo; 2020. https://doi.org/10.5281/ZENODO.4147317.

[197] Haileamlak A. Pandemics Will be More Frequent. Ethiop J Health Sci 2022;32:228. https://doi.org/10.4314/ejhs.v32i2.1.

[198] Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proceedings of the National Academy of Sciences 2021;118:e2105482118. https://doi.org/10.1073/pnas.2105482118.

[199] Ferrando C, Suarez-Sipmann F, Mellado-Artigas R, Hernández M, Gea A, Arruti E, et al. Clinical features, ventilatory management, and outcome of ARDS caused by COVID-19 are similar to other causes of ARDS. Intensive Care Med 2020;46:2200–11. https://doi.org/10.1007/s00134-020-06192-2.

[200] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Jair 2002;16:321–57. https://doi.org/10.1613/jair.953.

[201] Zhang P, Jia Y, Shang Y. Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks 2022;18:15501329221106935. https://doi.org/10.1177/15501329221106935.

[202] Kabir MF, Ludwig S. Classification of Breast Cancer Risk Factors Using Several Resampling Approaches. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, p. 1243–8. https://doi.org/10.1109/ICMLA.2018.00202.

[203] Meng C, Zhou L, Liu B. A Case Study in Credit Fraud Detection With SMOTE and XGBoost. J Phys: Conf Ser 2020;1601:052016. https://doi.org/10.1088/1742-6596/1601/5/052016.

[204] Yang J, Guan J. A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm. Information 2022;13:475. https://doi.org/10.3390/info13100475.

[205] Velarde G, Sudhir A, Deshmane S, Deshmunkh A, Sharma K, Joshi V. Evaluating XGBoost for Balanced and Imbalanced Data: Application to Fraud Detection 2023. http://arxiv.org/abs/2303.15218 (accessed January 26, 2024).

[206] Hajek P, Abedin MZ, Sivarajah U. Fraud Detection in Mobile Payment Systems using an XGBoost-based Framework. Inf Syst Front 2023;25:1985–2003. https://doi.org/10.1007/s10796-022-10346-6.

[207] Barker M, Chue Hong NP, Katz DS, Lamprecht A-L, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. Sci Data 2022;9:622. https://doi.org/10.1038/s41597-022-01710-x.

[208] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[209] Lamprecht A-L, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data Science 2020;3:37–59. https://doi.org/10.3233/DS-190026.

[210] Moonesinghe R, Khoury MJ, Janssens ACJW. Most Published Research Findings Are False—But a Little Replication Goes a Long Way. PLoS Med 2007;4:e28. https://doi.org/10.1371/journal.pmed.0040028.

[211] Simons DJ. The Value of Direct Replication. Perspect Psychol Sci 2014;9:76–80. https://doi.org/10.1177/1745691613514755.

[212] Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018;359:725–6. https://doi.org/10.1126/science.359.6377.725.

[213] Struck A, Loewe A, Achhammer E, Rack F, Bach F, Löffler F, et al. A Guide for Publishing, Using, and Licensing Research Software in Germany. Zenodo; 2020. https://doi.org/10.5281/ZENODO.4327147.

[214] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings : An updated review of related biases. Health Technology Assessment 2010;14:1–220. https://doi.org/10.3310/hta14080.

[215] Neill US. Publish or perish, but at what cost? J Clin Invest 2008;118:2368. https://doi.org/10.1172/JCI36371.

[216] Publish or perish. Nature 2010;467:252–2. https://doi.org/10.1038/467252a.

[217] APA Dictionary of Psychology n.d. https://dictionary.apa.org/ (accessed February 16, 2024).

[218] Jiménez R, Kuzak M, Alhamdoosh M, Barker M, Batut B, Borg M, et al. Four simple recommendations to encourage best practices in research software. F1000Research 2017;6. https://doi.org/10.12688/f1000research.11407.1.

[219] Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, et al. Machine learning for patient risk stratification: Standing on, or looking over, the shoulders of clinicians? Npj Digit Med 2021;4:1–6. https://doi.org/10.1038/s41746-021-00426-3.