Empowering advanced medical decision-making through machine learning in healthcare

Kumulative Dissertation

der Wirtschaftswissenschaftlichen Fakultät
der Universität Augsburg
zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)

Vorgelegt von
Milena Grieger, M.Sc.

16.11.2023

Erstgutachter: Prof. Dr. Jens O. Brunner
Zweitgutachter: Prof. Dr. Sebastian Schiffels

Vorsitzender der mündlichen Prüfung: Prof. Dr. Axel Tuma

List of Contributions

This thesis contains the following contributions submitted to or published in scientific journals. The specified categories relate to the journal ranking VHB-JOURQUAL3 of the Verband der Hochschullehrer für Betriebswirtschaft e.V. (2015). The order of the contributions corresponds to the order of print in this thesis.

Contribution 1: Grieger, M, Brunner, JO, Heller, AR, Bartenschlager, CC (2023). Scarce, scarcer, scarcest: Performance-flexible AI-based planning of elective surgeries for efficient and effective intensive care capacity management.
Status: Submitted on August 31, 2023, to *OR Spectrum*; Category A.

Contribution 2: Grieger, M, Shala, E, Schüller, M, Ebel, SS, Brunner, JO, Vehreshild, JJ, Erber, J, Hanses, F, Zabel, LT, Römmele, C, Shmygalev, S, Bartenschlager, CC (2023). *DENLU* and *leaky stanh*: customized activation functions targeting enhanced sensitivity with healthcare applications in binary classification**.**
Status: Close to submission.

Contribution 3: Bartenschlager, CC, Grieger, M, Erber, J, Neidel, T, Borgmann, S, Vehreschild, JJ, Steinbrecher, M, Rieg, S, Stecher, M, Dhillon, C, Ruethrich, MM, Jakob, CEM, Hower, M, Heller, AR, Vehreschild, M, Wyen, C, Messmann, H, Pipel, C, Brunner, JO, Hanses, F, Römmele, C (2023). Covid-19 triage in the emergency department 2.0: How analytics and AI transform a human-made algorithm for the prediction of clinical pathways.
Status: Published in *Health Care Management Science*; Category A.

Acknowledgments

I would like to express my deepest appreciation to my academic advisor, Prof. Dr. Jens O. Brunner, for your unwavering support, unshakeable confidence, and invaluable academic insights that have enriched my journey over the past three years. The countless discussions we've engaged in have been not only enlightening but also instrumental in shaping my academic growth. I also want to express my deep gratitude to Prof. Dr. Christina Bartenschlager for her profound and scientific support, which has greatly contributed to the depth and quality of my work. Additionally, I want to acknowledge and thank the remarkable individuals I had the privilege of working with. Your collaboration has made the academic environment not only intellectually stimulating but also an enjoyable place to spend time, both within and beyond the confines of our offices.

I could not have undertaken this journey without my family and friends. Your unwavering belief in me has consistently fueled my motivation and offered me with invaluable support. I want to convey my deep gratitude to you, Robin, particularly for being a constant source of both support and balance, even during the most stressful times, and for the joy you've brought into my life.

Perhaps we should all stop for a moment and focus not only on making our AI better and more successful but also on the benefit of humanity.

STEPHEN HAWKING

Contents

1 Introduction

Machine learning holds the potential to address challenges in the healthcare sector. This thesis investigates three distinct contributions within machine learning and assesses how applications and methodologies can positively influence medical decision-making.

1.1 Motivation

The healthcare system has grappled with significant challenges over the years, and the emergence of the Covid-19 pandemic has only underscored these issues. Notably, escalating costs (Bai et al. 2021), a shortage of nurses (Chan et al. 2013), and excessive bureaucratic expenses (Lorkowski et al. 2020) have resulted in overburdened healthcare professionals and consequently dissatisfied patients. Figure 1 depicts the issue of rising costs within the German healthcare system, along with the substantial increase in the number of people in need of care. This increase surpasses the growth rate of healthcare personnel, with 2011 as the reference year. To sustain and enhance patient care, multiple approaches can be considered. One way is the usage and expansion of machine learning techniques to alleviate the workload of clinical staff and supporting the decision-making process (Haug and Drazen 2023).

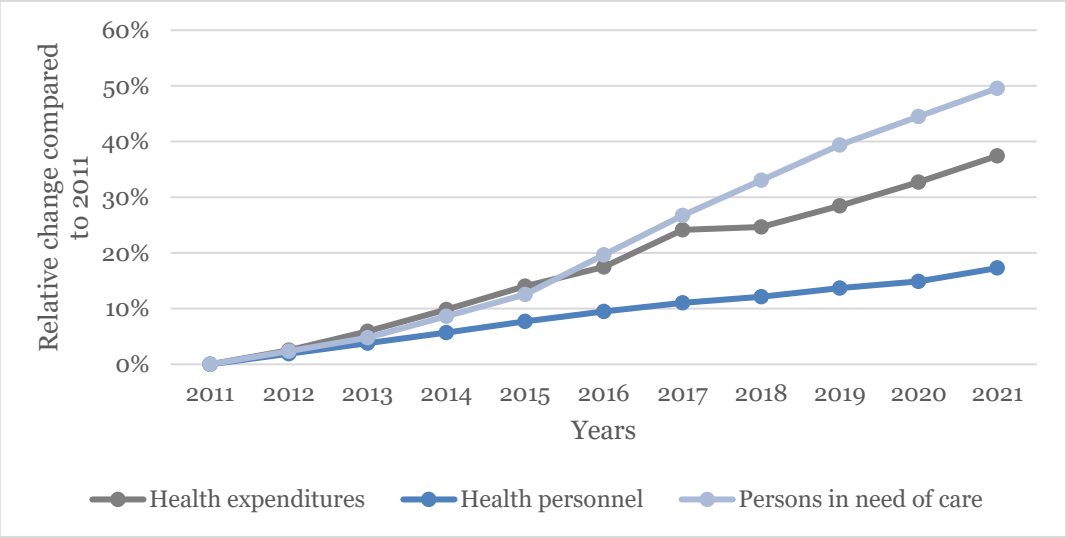

Figure 1: Relative change in health expenditures, health personnel and persons in need of care in Germany from 2011 (reference year) to 2021 (based on Federal Statistical Office Germany (2023a, 2023b, 2023c)

Medical decision-making impacts everyone at some point, whether in the role of a patient or a healthcare professional. Its significance is further underscored by its potential application across operational, tactical, and strategic levels. Medical decision-makers often grapple with intricate dilemmas, where information may be limited, yet critical decisions must be made (Masic 2022). A physician in the emergency room, for example, faces the critical choice of whether to admit a patient to the intensive care unit (ICU). Various computer tools can aid in making such decisions (Awaysheh et al. 2019). One promising avenue involves harnessing machine learning support to mitigate uncertainty and enhance accuracy.

1.2 Introduction to machine learning

Machine learning has become a ubiquitous term in recent years and is already making substantial contributions across various domains. Its journey began in the 1950s with landmarks like Rosenblatt's Mark-1 perceptron (Rosenblatt 1958) and Samuel's Game of Checkers (Samuel 1959). Today, machine learning applications have become a norm, finding utility in diverse fields such as supply chain management, mobility solutions, and marketing strategies. These applications range from assisting in management decisions to the development of novel machine learning methodologies.

Machine learning falls within the broader field of artificial intelligence (AI), with deep learning being a subset of it. Machine learning, in turn, can be categorized into the three main domains supervised learning, unsupervised learning, and reinforcement learning (Joshi 2020). Supervised learning involves the utilization of labeled data during the training process, enabling the model to discern relevant patterns and associations (Nasteski 2017). Unsupervised learning aims to uncover underlying patterns within data without the presence of explicit output labels (Dike et al. 2018). Lastly, reinforcement learning relies on a trial-and-error approach, learning from its own actions and experiences based on feedback (Sutton and Barto 2018). Figure 2 illustrates the categories of machine learning models and thematically organizes the three contributions to this thesis. Further explanation is provided in section 2.

This study predominantly centers on supervised learning tasks, which can be further divided into regression and classification problems, with a primary focus on binary classification. Furthermore, the selection of algorithms is crucial, including options such as deep neural networks, decision trees, random forests, extreme gradient boosting, and logistic regression.

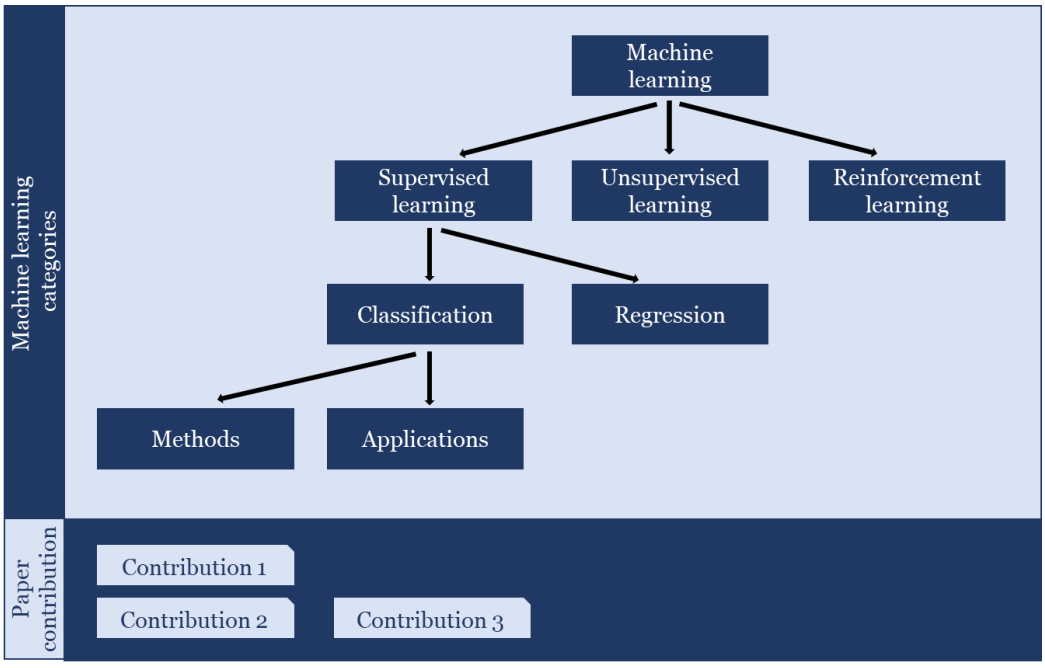

Figure 2: Machine learning categories with an emphasis on supervised learning

In the realm of machine learning, exploring new frontiers involves not only innovation in terms of algorithms, but also transformation in the learning process. This transformation encompasses the adjustment of activation functions (AFs), responsible for converting input into output through weighted sums (Sharma et al. 2020), as well as the design of loss functions, which quantify the error between predictions and real values, aiming for minimization (Wang et al. 2022). These choices play a pivotal role in shaping the performance metrics used to evaluate machine learning algorithms, with common measures including accuracy, sensitivity, and specificity. Accuracy gauges the overall correctness of predictions, while sensitivity emphasizes correct positive predictions (true positives), and specificity focuses on correct negative predictions (true negatives) (Sokolova and Lapalme 2009). Although accuracy often takes precedence in general machine learning applications, there are scenarios where the emphasis shifts to a specific class, such as the positive class, requiring a heightened sensitivity. Achieving this objective may necessitate methodical adjustments to AFs and loss functions.

1.3 Machine learning in healthcare

Machine learning technology is gaining also substantial traction in the field of medicine. Researchers are actively exploring applications like image recognition in MRI scans (Razzak et al. 2018), the prediction of diagnoses (Iqbal et al. 2021), patient trajectories (Pham et al. 2017), and resource capacities (Quiroz-Juárez et al. 2021). Machine learning's involvement in the healthcare industry is expanding to encompass the corporate sector as well. Prominent companies like Microsoft and Pfizer have adopted machine learning processes, particularly in the field of cancer detection. Additionally, a surge of startups, including MedInReal for smart health records (MedInReal 2023) and Cancer Center specializing in image analysis (Cancer Center 2023), has emerged in recent years, aiming to alleviate the burdens on the healthcare system. According to Qayyum et al. (2021), machine learning in healthcare encompasses the four primary application areas prognosis, diagnosis, treatment and clinical workflow. Prognosis involves tracking the development of diseases, including symptoms and complications. Diagnosis finds its main

applications in electronic health records and image analysis. Machine learning supports treatment through tasks like annotating reports using natural language processing and real-time health monitoring. The final category, clinical workflows, comprises subcategories such as disease prediction and the analysis of clinical time-series data (Qayyum et al. 2021). The research papers in this thesis primarily focus on prognosis and clinical workflows, particularly in the context of the ICU within hospital settings.

Given the sensitive nature of healthcare data, considerations extend beyond technical aspects. Topics such as explainable AI, which enhances user understanding of predictions (Gunning et al. 2019), and ethical concerns (Rigby 2019), like the appropriateness of predicting patient mortality, are integral to this domain. Moreover, medical applications often demand specialized machine learning algorithms. For instance, in binary classification scenarios, one class may hold particular importance, as seen in the scarcity of ICU beds compared to regular ward beds, necessitating a heightened focus on sensitivity. This can be achieved through customized loss functions and AFs (see section 2.1 and section 2.2). Furthermore, there are uncharted territories in medicine where machine learning holds the potential to unveil valuable insights, offering promising avenues for future exploration (See section 2.3).

1.4 Organization of this thesis

In this dissertation, advanced medical decision-making is investigated and addressed with three contributions to precisely answer the following research questions:

1. Can a loss function be harnessed to empower decision-makers in the flexible planning of capacities within ICUs?

2. Can customized AFs enhance the performance of sensitivity-based binary classification?

3. Can the integration of analytics and machine learning methods lead to improvements in Covid-19 triage for clinical pathways, while ensuring the explainability of the algorithms?

The subsequent sections of this dissertation are meticulously organized as follows. Section 2 offers a concise overview of the three distinct papers that collectively constitute the contributions of this research. Section 3 delves into a comprehensive discussion of the papers, providing in-depth elaborations on the research questions above. Section 4 encompasses the culmination of this work, presenting the concluding remarks, and paving the way for future research endeavors. The complete versions of the unpublished contributions and a link to the published contribution can be found in the Appendix.

2 Summary of the contributions

This thesis offers several valuable contributions to the current literature, showcasing the potential of machine learning to empower advanced decision-making in healthcare. The following section provides a comprehensive summary of these contributions, with the entire contributions available in the Appendix. It is important to note that the sequence of these contributions does not follow the chronological order of submission to scientific journals. Instead, the discussion begins with an exploration of supervised learning methods designed to improve sensitivity, concluding with an examination of the application of various models (see Figure 2). The first paper introduces a performance-flexible AI-based planning approach that incorporates a loss function and simulation to cater to the diverse needs of decision-makers. Following that, the subsequent paper delves into sensitivity-centered binary classification, employing customized AFs and dealing with diverse hospital data sets. Finally, the last contribution centers on the application of supervised learning, encompassing various approaches with varying levels of explanatory power, all geared toward improving the triage process for Covid-19 patients.

2.1 Scarce, scarcer, scarcest: Performance-flexible AI-based planning of elective surgeries for efficient and effective intensive care capacity management

Grieger et al. (2023a) examine how a loss function affects sensitivity, specificity, and consequently, the predictive capacity for distinguishing between ICU and non-ICU cases, aimed at aiding medical decision-makers. This research falls under the category of the methodology of supervised learning (see Figure 2). The complete paper can be accessed in its entirety in Appendix A and has been submitted to *OR Spectrum*, a journal ranked in category A according to the VHB-JOURQUAL3 ranking system (Verband der Hochschullehrer für Betriebswirtschaft e.V. 2015).

Motivation. In the planning of elective surgeries, the availability of ICU and operating room (OR) capacity plays a major role, as both resources are typically scarce within a hospital setting and necessitate integrated planning. Decision-makers tasked with determining whether a patient requires ICU treatment can benefit from the support of machine learning. However, mere accuracy, as commonly emphasized in the existing literature, is insufficient, especially when both OR and ICU capacity are constrained. A flexible approach that prioritizes either the accurate prediction of ICU (sensitivity) or non-ICU treatment (specificity) is essential. While prior research has explored various loss functions to balance data, the focus has not been on achieving flexible performance. Another method known as thresholding pursues flexibility but is applied post-training.

Methods. To address this gap, this paper introduces a performance-flexible AI-based planning approach for predicting the need for ICU treatment after elective surgery. The approach incorporates a performance-flexible loss function within a machine learning framework, including subsequent simulation of ICU occupancy. The problem at hand involves supervised learning for binary classification of patients. The *performance-flexible binary cross-entropy (PFBCE)* loss function introduced in this work is based on *cost-sensitive binary cross-entropy* loss function (Aurelio et al. 2019) but centers

on flexible performance rather than data balance. The flexibility of *PFBCE* can be influenced by adjusting the weight of the positive class and its convex weight of the negative class. A weight of 0.5 implies equal importance to both classes, while a weight greater than 0.5 prioritizes sensitivity and a weight smaller than 0.5 prioritizes specificity. Additionally, a normalized version of the loss function, known as *normalized PFBCE (NPBCE)*, is introduced to allow users to choose between *PFBCE* and *NFBCE* based on the conditions via a hyperparameter.

Application. To evaluate the approach, we used a data set from the University Hospital of Augsburg in Germany, spanning over 26,600 elective surgeries conducted between 2017 and 2021. We employ logistic regression and deep neural networks as machine learning techniques, involving various steps such as data preparation, hyperparameter tuning, data balancing, and 5-fold cross-validation. The simulation component of the study aims to assess different weightings within the performance-flexible AI-based planning approach under varying scenarios. These scenarios involve different scarcity patterns in OR and ICU resources and categorize patients into cohorts based on their ASA-scores, distinguishing between patients with low, average, and high ASA-scores. The ASA-score is a widely used scoring system to categorize patients based on their physical condition (Saklad 1941). Monte Carlo simulations with 1,000 runs are conducted for each weight, scarcity pattern, and patient cohort. The key performance metric calculated for each simulation run is the ratio of realized ICU patients to planned ICU capacity, with an average value calculated for each weight, scarcity pattern, and patient cohort. A ratio of 1 indicates that realized ICU utilization matches the planned capacity, while a ratio greater than 1 implies higher demand than planned and a ratio less than 1 suggests lower occupancy than planned. The generalizability of this approach is underscored by its reliance on widely available hospital data.

Results. Comparative analysis demonstrates that the performance-flexible AI-based planning approach can dynamically prioritize specific labels in

binary classification while maintaining high accuracy. Consequently, the ratio of realized demand to planned ICU capacity remains close to 1 across diverse simulation scenarios, encompassing resource scarcity and patient cohorts. This stands in contrast to traditional machine learning approaches, which are suitable primarily for scenarios involving relatively healthy patients. The presented approach holds promise in assisting hospital decision-makers while offering planning flexibility.

2.2 *DENLU* and *leaky stanh*: customized activation functions targeting enhanced sensitivity with healthcare applications in binary classification

Grieger et al. (2023b) investigate the potential of customized AFs in healthcare to enhance sensitivity. Their exploration encompasses various aspects, including the Universal Approximation Theorem (UAT) and data structure considerations. This research falls under the category of the methodology of supervised learning (see Figure 2). The complete paper can be accessed in its entirety in Appendix B and is close to submission.

Motivation. In healthcare, binary classification of patients is a common practice. This involves categorizing patients into distinct groups, such as determining positive or negative diagnoses, inpatient or outpatient status, or the need for ICU versus non-ICU. These classifications play a crucial role in resource planning for effective and efficient medical care. Unlike many machine learning approaches that assume equal importance for both binary classes, healthcare often prioritizes a specific class, such as ICU. Moreover, the significance of data structure and data quality is frequently overlooked when employing machine learning models in healthcare applications. Additionally, there is a noticeable disparity between the extensive theoretical groundwork in machine learning and the limited development of new methods. Consequently, only a few AFs are available for integration into algorithms, and even these are only partially aligned with the principles of the UAT.

Methods. To address these unresolved issues, this work introduces two customized AFs, the *Double Exponential Non-Linear Unit* (*DENLU*) and the *Leaky Scaled Hyperbolic Tangent* (*leaky stanh*). These functions are built upon the well-established AFs *Exponential Linear Unit* (*ELU*) and *Scaled Hyperbolic Tangent* (*stanh*) and have been modified to align as closely as possible with the principles of the UAT. UAT serves as a blueprint for crafting an AF and comprises five core principles (Sodhi and Chandra 2014). The primary objective of *DENLU* and *leaky stanh* is to mitigate the limitations associated with existing functions, including issues like vanishing gradients, inflexible shapes, and training instability, while simultaneously enhancing sensitivity in comparison to known AFs. Vanishing gradients refer to the phenomenon where gradients diminish progressively, approaching zero as the independent variable's absolute values increase (Li et al. 2014). *DENLU* takes the form of a sigmoidal, locally quadratic function and adheres to all five UAT principles. It is derivable for all values, the output is within the range of -1 to 1, and includes an adjustable parameter for fine-tuning of the range. On the other hand, *leaky stanh* is a sigmoidal and piecewise linear function, satisfying four out of five UAT principles. This AF primarily aims to tackle the vanishing gradient issue. Both of these adapted AFs offer flexibility in shaping and ensure training stability.

Application. The two AFs are tested on a total of four binary classification data sets, encompassing three real-world healthcare and one simulated data set. Among these, three data sets exhibit heterogeneous features, while the remaining data set comprises homogeneous features. The evaluation process includes the utilization of deep neural networks, 10-fold cross-validation, hyperparameter tuning, and various data preprocessing steps tailored to each specific data set. Furthermore, the performance of *DENLU* and *leaky stanh* is contrasted with four established AFs.

Results. The findings reveal that *DENLU*, offering enhanced flexibility compared to *sigmoid* with improved shaping capabilities, and *leaky stanh*, serving as an alternative to *stanh* while mitigating vanishing gradient issues,

boast theoretical advantages aligned with the UAT. Moreover, these two AFs deliver superior results in terms of sensitivity, with improvements of up to 17.7 percentage points and in the area under the curve (AUC) showing enhancements of up to 7.6 percentage points. Notably, these advantages are particularly pronounced in the context of heterogeneous data sets, emphasizing that the primary application of these two AFs should be directed toward such data sets. The outcomes illustrate that decision-makers can enhance their decision-making support, particularly within specific classes like ICU treatment, by integrating AFs such as *DENLU* or *leaky stanh*.

2.3 Covid-19 triage in the emergency department 2.0: How analytics and AI transform a human-made algorithm for the prediction of clinical pathways

Bartenschlager et al. (2023) investigate the viability of analytical and AI-based methods in the triage of Covid-19 patients, with a particular emphasis on the explainability of these algorithms. This research falls under the category of the applications of supervised learning (see Figure 2) and has been published in *Health Care Management Science*, a journal ranked in category A according to the VHB-JOURQUAL3 ranking system (Verband der Hochschullehrer für Betriebswirtschaft e.V. 2015).

Motivation. The Covid-19 pandemic has strained the capacities of numerous hospitals, shifting the spotlight onto patient triage, a topic that has sparked extensive discussions from various perspectives, including ethical considerations. The concept of triage has long been employed in settings like emergency rooms to prioritize treatment based on urgency (FitzGerald et al. 2010) or in mass casualty incidents to maximize lives saved (Neidel et al. 2017). In the context of Covid-19, triage merges these principles and encompasses diverse factors, such as treatment urgency, disease severity, and the categorization of clinical pathways. Categorizing patients based on clinical pathways determines their placement in ward, ICU, discharge, palliative care unit (PCU), and initiates in the emergency department (ED). Despite its critical implications for patient care and capacity planning, this classification is

notably underrepresented in the existing literature. In contrast, areas like Covid-19 diagnosis have seen an abundance of publications with a primary focus on AI. However, practical application frequently lacks consideration of real-world implementation, transparency, comprehensive databases, and validation processes (Wynants et al. 2020).

Methods. This work focuses on evaluating and scrutinizing the performance of various triage algorithms, both analytical and AI-based, with a specific focus on their explainability and ethical considerations. The foundation for the Covid-19 triage is a base triage algorithm (TA) initially proposed by Pin et al. (2020) for patient classification concerning clinical pathways. This algorithm has served as a guideline for EDs in Germany, as recommended by the German Society for Interdisciplinary Emergency and Acute Medicine. The TA is designed as a straightforward, easily comprehensible decision tree, and it has already been applied at institutions such as the University Hospital of Augsburg. It categorizes patients into the three labels ward, ICU, and discharge. Building upon this base, the study seeks to extend the TA with a data-driven approach, referred to as the extended triage algorithm (TAE), which now incorporates the fourth label, PCU. Both TA and TAE can be classified as white-box. In addition to these analytical approaches, AI-based triage algorithms, such as Multi-Layer Perceptron (MLP), Random Forest (RF), and Extreme Gradient Boosting (XGB), are considered. It is important to note that these AI models are often viewed as black-boxes, lacking transparency. The study also explores an integrated approach, combining both analytical and AI-based methods (integrated triage algorithm (ITA)) in a two-step process. Prior to a physician's real triage, which employs a data-guided decision tree, an initial pre-triage phase is executed using AI. This combined approach aims to evaluate a human-AI interactive algorithm, with both autonomous black-box and white-box components. All algorithms, except TA and ITA, are assessed for both three-label (ward, ICU, discharge) and four-label (ward, ICU, discharge, PCU) classification to address ethical considerations. Furthermore, the study investigates various data preparation methods, including the impact of different imputation techniques for handling

missing values and diverse strategies for summarizing comorbidities, to assess their influence on algorithm performance.

Application. To assess the performance of the analytical and AI-based triage algorithms, we employ a data set obtained from the Lean European Open Survey on Covid-19 Patients. This data set encompasses information from 4,310 Covid-19 patients recorded until January 2021, spanning the first and second pandemic waves in Europe.

Results. The findings indicate a substantial improvement in the performance of the TAE in comparison with the TA. Both TA and TAE are characterized by their straightforward, easy-to-interpret decision trees. Comparatively, the AI-based algorithms (MLP, RF, XGB) and the ITA exhibit similar performance but significantly outperform both TA and TAE. ITA offers a significant advantage with its heightened sensitivity, particularly in the context of ICU prediction. Additionally, it provides a level of explainability by combining machine and human decision-making, rendering it a preferred option. Explainability is particularly crucial since the algorithm directly impacts patients and medical staff in the ED. The results are not substantially affected by variations in data preparation methods. Furthermore, considering the inclusion of the fourth label, PCU, is not advisable from both an ethical and data-driven perspective. These findings hold significant implications, especially in the context of planning for highly occupied ICUs.

3 Discussion of the contributions

This thesis is devoted to advanced medical decision-making through machine learning in healthcare. The next three subsections discuss how the three contributions introduced address the following research questions.

1. Can a loss function be harnessed to empower decision-makers in the flexible planning of capacities within ICUs?
2. Can customized AFs enhance the performance of sensitivity-based binary classification?
3. Can the integration of analytics and machine learning methods lead to improvements in Covid-19 triage for clinical pathways, while ensuring the explainability of the algorithms?

3.1 Can a loss function be harnessed to empower decision-makers in the flexible planning of capacities within ICUs?

Grieger et al. (2023a) have introduced a performance-flexible AI-based planning approach designed to enhance the efficiency and effectiveness of ICU capacity management. In this section, a discussion about whether a loss function integrated into an algorithm can provide decision support in the flexible planning of ICUs is provided.

The performance-flexible AI-based planning approach, as discussed previously, comprises both a machine learning model and a simulation. To address the research question, our primary focus will be on the machine learning model, while the simulation is used to analyze the results.

The machine learning model applied in this study incorporates a deep neural network and logistic regression, leveraging a novel loss function, the *(N)PFBCE*. This loss function is engineered to deliver flexible performance by adjusting weights for the positive and negative classes. When implemented in the context of ICU capacity management, this convex approach allows the prioritization of the ICU or non-ICU class, particularly crucial during periods of limited ICU capacity.

The performance-flexible AI-based approach is evaluated using a data set encompassing over 26,000 elective patients. The results illustrate that an increased focus on the ICU class leads to heightened sensitivity without significant losses in specificity and accuracy, and vice versa. For instance, with a weight of 0.8, indicating a strong emphasis on correctly predicting the ICU class, *NPFBCE* and the deep neural network achieve a sensitivity of 98.29 %, a specificity of 76.78 %, and an accuracy of 87.53 %. The simulation delves into various scenarios, accounting for differing OR and ICU capacities, and categorizes patients into three distinct health groups. The findings underscore the growing significance of accurately predicting the ICU class, particularly for more critically ill patients, as this is the sole means to prevent ICU overload. Consequently, medical decision-makers can adjust their predictive focus in alignment with capacity constraints, aiding their decision-making process.

In summary, the utilization of *PFBCE* in binary classification results in improved capacity planning, offering enhanced decision support within ICUs.

3.2 Can customized AFs enhance the performance of sensitivity-based binary classification?

Grieger et al. (2023b) introduce customized AFs designed to prioritize sensitivity in binary classification of patients. This section discusses the influence of customized AFs on the sensitivity of a binary classification.

The machine learning approach outlined in the paper encompasses a neural network and two customized AFs, *DENLU* and *leaky stanh*. These two AFs demonstrate intriguing theoretical properties, drawing from the UAT and

offering flexibility in their S-shaped curves, while partially mitigating the vanishing gradient issue. *DENLU* and *leaky stanh* are evaluated using three real-world healthcare data sets and one simulated data set, all characterized by binary classifications. The findings indicate that these customized AFs are beneficial for handling heterogeneous data sets, which includes three out of the four data sets analyzed. In these cases, both AFs yield comparable results to well-established AFs in terms of accuracy. Notably, performance metrics such as AUC and sensitivity hold special significance in healthcare, as accurate predictions in the positive class have a profound impact on medical care and effective planning in hospitals. In terms of sensitivity, both *leaky stanh* and *DENLU* outperform other AFs, demonstrating improvements of up to 17.7 percentage points. This enhancement empowers medical decision-makers to make informed choices within specific classes, such as ICU.

In conclusion, the adoption of customized AFs, grounded in sound theoretical foundations, significantly elevates sensitivity in the binary classification of heterogeneous data sets.

3.3 Can the integration of analytics and machine learning methods lead to improvements in Covid-19 triage for clinical pathways, while ensuring the explainability of the algorithms?

Bartenschlager et al. (2023) conducted an assessment of an existing, an analytic, and several machine learning algorithms for triage of Covid-19 patients. This section examines whether a combination of analytic and machine learning methods can enhance Covid-19 triage while preserving transparency through the analytic component.

Beyond the standalone use of purely analytic or machine learning models, the integrated human-AI algorithm, denoted as ITA, is also considered. One notable advantage of ITA is that it avoids the sole deployment of a black-box approach for predictions, as it enhances transparency through the analytical component. This aspect holds particular importance in the context of triage,

where numerous ethical considerations come into play. The algorithms are applied to an existing triage system with a data set comprising over 4,000 Covid-19 patients. The results reveal that ITA exhibits performance similar to other AI-based models, with all AI methods significantly outperforming the purely analytic models. However, ITA stands out as a combination of analytical and AI elements, providing substantial benefits owing to its partial transparency.

In summary, the integration of analytic and machine learning methods results in enhanced Covid-19 triage capabilities while maintaining transparency through partial explainability.

4 Conclusion

This thesis outlines the opportunities for empowering advanced medical decision-making through machine learning in healthcare. It initiates by highlighting the importance of medical decision-making in addressing healthcare challenges and introduces the concept of machine learning. Subsequently, it elucidates the role of contributions within the established frameworks and provides practical applications of machine learning models. The first and second contribution, focusing on loss functions and AFs, are categorized within the methodological domain of supervised learning and serve to aid medical decision-makers in capacity-constrained scenarios. The third contribution delves into the application aspect of supervised learning by introducing and testing various algorithms for Covid-19 triage, encompassing explainability, ethical considerations, and an integrated triage approach combining black- and white-box.

Within the area of supervised learning, all three approaches empower advanced medical decision-making. The first paper introduces a performance-flexible AI-based planning approach for elective surgeries, optimizing ICU capacity management. In addition to presenting and applying a novel loss function for machine learning, it involves simulations to address diverse capacity challenges. The outcomes demonstrate the adaptability of the performance-flexible AI-based planning approach, enabling decision-makers to flexibly prioritize specific binary classification labels and enhance their planning capabilities. The second paper explores two modified AFs rooted in the UAT, aiming to elevate sensitivity in binary classification.

These AFs exhibit promising properties, including S-shaped flexibility and mitigated vanishing gradient issues from a theoretical standpoint. The results reveal significantly improved sensitivity of the modified AFs, particularly for heterogeneous data sets, facilitating more accurate decision-making. The third paper assesses the impact of AI and analytics on an existing Covid-19 triage algorithm. It considers an established algorithm employed in German hospitals, an analytical extension, three machine learning models, and an integrated approach combining analytics and machine learning. The findings indicate that algorithms incorporating machine learning substantially outperform others, with the integrated approach being particularly advantageous due to its explanatory power, providing decision-makers with additional insights. All three contributions collectively provide added value not only for the topic of AI itself but also in the context of humanity, owing to their focus on specific situations within the healthcare sector.

These three contributions mark the inception of further research opportunities in the field of machine learning for medical decision-making. While the contributions predominantly center on supervised learning with pre-existing data, it prompts the intriguing question of how unsupervised learning and reinforcement learning can influence medical decision-making in situations where data availability is limited. Given the substantial disparity between machine learning applications and novel methods in healthcare, there is a clear need for additional methodologies that can bolster healthcare practices. Furthermore, the evaluation of these contributions is focused on specific data sets, opening up avenues for research using different data sets that could further validate the positive impact of the presented work. Investigations with data from diverse countries, distinct from Germany, present an intriguing avenue for exploration. Additionally, the methodological aspects of these contributions may have relevance beyond healthcare and could be applied to various other domains.

References

Aurelio YS, Almeida GM de, Castro CL de, Braga AP (2019) Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. Neural Process Lett 50:1937–1949. https://doi.org/10.1007/s11063-018-09977-1

Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL (2019) Review of Medical Decision Support and Machine-Learning Methods. Veterinary Pathology 56:512–525. https://doi.org/10.1177/0300985819829524

Bai J, Fügener A, Gönsch J, Brunner JO, Blobner M (2021) Managing admission and discharge processes in intensive care units. Health Care Manag Sci 24:666–685. https://doi.org/10.1007/s10729-021-09560-6

Bartenschlager CC, Grieger M, Erber J, Neidel T, Borgmann S, Vehreschild JJ, Steinbrecher M, Rieg S, Stecher M, Dhillon C, Ruethrich MM, Jakob CEM, Hower M, Heller AR, Vehreschild M, Wyen C, Messmann H, Piepel C, Brunner JO, Hanses F, Römmele C (2023) Covid-19 triage in the emergency department 2.0: how analytics and AI transform a human-made algorithm for the prediction of clinical pathways. Health Care Manag Sci 26:412–429. https://doi.org/10.1007/s10729-023-09647-2

Cancer Center (2023) Cancer Center - Platform in Oncology and Pathology. https://cancercenter.ai/. Accessed 9 October 2023

Chan ZCY, Tam WS, Lung MKY, Wong WY, Chau CW (2013) A systematic literature review of nurse shortage and the intention to leave. Journal of Nursing Management 21:605–613. https://doi.org/10.1111/j.1365-2834.2012.01437.x

Dike HU, Zhou Y, Deveerasetty KK, Wu Q (2018) Unsupervised Learning Based On Artificial Neural Network: A Review. In: 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS). IEEE

Federal Statistical Office Germany (2023a) Health expenditures in Germany as share of GDP and in millions of Euro. https://www.gbe-bund.de/gbe/pkg_isgbe5.prc_menu_olap?p_uid=gastd&p_aid=18559448&p_sprache=E&p_help=2&p_indnr=522&p_indsp=&p_ityp=H&p_fid=. Accessed 9 October 2023

Federal Statistical Office Germany (2023b) Health personnel: Germany, years, facilities, sex. https://www-genesis.destatis.de/genesis/online?sequenz=tabelleErgebnis&selectionname=23621-0001&language=en#abreadcrumb. Accessed 9 October 2023

Federal Statistical Office Germany (2023c) Pflegebedürftige (Anzahl und Quote). https://www.gbe-bund.de/gbe/pkg_isgbe5.prc_menu_olap?p_uid=gast&p_aid=27610956&p_sprache=D&p_help=0&p_indnr=510&p_indsp=138&p_ityp=H&p_fid=. Accessed 9 October 2023

FitzGerald G, Jelinek GA, Scott D, Gerdtz MF (2010) Emergency department triage revisited. Emergency Medicine Journal 27:86–92. https://doi.org/10.1136/emj.2009.077081

Grieger M, Brunner JO, Heller AR, Bartenschlager CC (2023a) Scarce, scarcer, scarcest: Performance-flexible AI-based planning of elective surgeries for efficient and effective intensive care capacity management. Working Paper, University of Augsburg

Grieger M, Shala E, Schüller M, Ebel SS, Brunner JO, Vehreshild JJ, Erber J, Hanses F, Zabel LT, Römmele C, Shmygalev S, Bartenschlager CC (2023b) Sensitivity-centered binary classification of patients using customized activation functions and machine learning. A case study. Working Paper, University of Augsburg

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI-Explainable artificial intelligence. Science Robotics 4. https://doi.org/10.1126/scirobotics.aay7120

Haug CJ, Drazen JM (2023) Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med 388:1201–1208. https://doi.org/10.1056/NEJMra2302038

Iqbal MJ, Javed Z, Sadia H, Qureshi IA, Irshad A, Ahmed R, Malik K, Raza S, Abbas A, Pezzani R, Sharifi-Rad J (2021) Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int 21:270. https://doi.org/10.1186/s12935-021-01981-1

Joshi AV (ed) (2020) Machine learning and artificial intelligence. Springer, Cham

Li J-C, Ng WWY, Yeung DS, Chan PPK (2014) Bi-firing deep neural networks. Int. J. Mach. Learn. & Cyber. 5:73–83. https://doi.org/10.1007/s13042-013-0198-9

Lorkowski J, Maciejowska-Wilcock I, Pokorski M (2020) Overload of Medical Documentation: A Disincentive for Healthcare Professionals. In: Medical Research and Innovation. Springer, Cham, pp 1–10

Masic I (2022) Medical Decision Making - an Overview. Acta Inform Med 30:230–235. https://doi.org/10.5455/aim.2022.30.230-235

MedInReal (2023) MedInReal - The AI virtual assistant for doctors. http://www.medinreal.com/. Accessed 9 October 2023

Nasteski V (2017) An overview of the supervised machine learning methods, vol 4

Neidel T, Salvador N, Heller AR (2017) Impact of systolic blood pressure limits on the diagnostic value of triage algorithms. Scand J Trauma Resusc Emerg Med 25:118. https://doi.org/10.1186/s13049-017-0461-2

Pham T, Tran T, Phung D, Venkatesh S (2017) Predicting healthcare trajectories from medical records: A deep learning approach. J Biomed Inform 69:218–229. https://doi.org/10.1016/j.jbi.2017.04.001

Pin M, Künstler C, Dodt C, Jerusalem K (2020) Behandlung Covid-19 Verdachtsfälle in der Notaufnahme, DGINA Notfallcampus V1.03, 2020, modified version according to K. Weber, Klinikum Kassel: COVID-19 Abklärungsalgorithmus Erwachsene (according to UCSF COVID-19 ID Clinical Working Group) and Zhang et al.: Therapeutic and triage strategies for 2019 novel coronavirus disease in fever clinics. Lanc Resp Med 8:e11-e12

Qayyum A, Qadir J, Bilal M, Al-Fuqaha A (2021) Secure and Robust Machine Learning for Healthcare: A Survey. IEEE Rev Biomed Eng 14:156–180. https://doi.org/10.1109/RBME.2020.3013489

Quiroz-Juárez MA, Torres-Gómez A, Hoyo-Ulloa I, León-Montiel RdJ, U'Ren AB (2021) Identification of high-risk COVID-19 patients using machine learning. PLOS ONE 16:e0257234. https://doi.org/10.1371/journal.pone.0257234

Razzak MI, Naz S, Zaib A (2018) Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps: Automation of Decision Making,:323–350

Rigby MJ (2019) Ethical Dimensions of Using Artificial Intelligence in Health Care. AMA Journal of Ethics 21:E121-124. https://doi.org/10.1001/amajethics.2019.121

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408. https://doi.org/10.1037/h0042519

Saklad M (1941) Grading of patients for surgical procedures. Anesthesiology 2:281–284. https://doi.org/10.1097/00000542-194105000-00004

Samuel AL (1959) Some Studies in Machine Learning Using the Game of Checkers. IBM J. Res. & Dev. 3:210–229. https://doi.org/10.1147/rd.33.0210

Sharma S, Sharma S, Athaiya A (2020) Activation functions in neural networks. IJEAST 04:310–316. https://doi.org/10.33564/ijeast.2020.v04i12.054

Sodhi SS, Chandra P (2014) Bi-modal derivative activation function for sigmoidal feedforward networks. Neurocomputing 143:182–196. https://doi.org/10.1016/j.neucom.2014.06.007

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Information Processing & Management 45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sutton RS, Barto A (2018) Reinforcement learning: An introduction. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, London, England

Verband der Hochschullehrer für Betriebswirtschaft e.V. (2015) Liste der Fachzeitschriften in VHB-JOURQUAL3. https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/gesamtliste. Accessed 10 October 2023

Wang Q, Ma Y, Zhao K, Tian Y (2022) A Comprehensive Survey of Loss Functions in Machine Learning. Ann. Data. Sci. 9:187–212. https://doi.org/10.1007/s40745-020-00253-5

Wynants L, van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Dahly DL, Damen JAA, Debray TPA, Jong VMT de, Vos M de, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Heus P, Kammer M, Kreuzberger N, Lohmann A, Luijken K, Ma J, Martin GP, McLernon DJ, Andaur Navarro CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, Tzoulaki I, van Kuijk SMJ, van Bussel B, van der Horst ICC, van Royen FS, Verbakel JY,

Wallisch C, Wilkinson J, Wolff R, Hooft L, Moons KGM, van Smeden M (2020) Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369:m1328. https://doi.org/10.1136/bmj.m1328

Appendix A: Performance-flexible AI

Grieger, M, Brunner, JO, Heller, AR, Bartenschlager, CC (2023). Scarce, scarcer, scarcest: Performance-flexible AI-based planning of elective surgeries for efficient and effective intensive care capacity management.

Status: Submitted on August 31, 2023, to *OR Spectrum*; Category A.

Original Research Paper:

Scarce, scarcer, scarcest: Performance-flexible AI-based planning of elective surgeries for efficient and effective intensive care capacity management

Milena Grieger[1], Jens O. Brunner[1,2], Axel R. Heller[3], Christina C. Bartenschlager[3,4]

[1]Health Care Operations/Health Information Management, Faculty of Business and Economics, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[2]Department of Technology, Management, and Economics, Technical University of Denmark
[3]Anaesthesiology and Operative Intensive Care, University Hospital of Augsburg, Germany
[4]Professor of Applied Data Science in Health Care, Nürnberg School of Health, Ohm University of Applied Sciences Nuremberg, Germany

Correspondence:
Jens Brunner, jotbr@dtu.dk

Submission: August 2023

Acknowledgements: We thank the German Federal Ministry of Education and Research for funding of the KISIK project. Grant Numbers 16SV9029 (University of Augsburg), 16SV9030 (University Hospital of Augsburg)

Financial support: This study was (partially) funded by the German Federal Ministry of Education and Research.

Conflicts of interest: The authors Bartenschlager, Brunner and Heller declare funding by the German Federal Ministry of Education and Research. The author Grieger declares no conflicts of interest.

Data availability statement: The data set generated during the current study is not publicly available as it contains exclusively sensitive healthcare data. Information on how to obtain it and reproduce the analysis is available from the corresponding author on request.

Original Research Paper:

Scarce, scarcer, scarcest: Performance-flexible AI-based planning of elective surgeries for efficient and effective intensive care capacity management

Abstract. Operating room and intensive care unit (ICU) capacities belong to the scarcest resources in hospitals and strongly depend on each other. When planning elective surgeries, it is therefore important to consider both resources in an integrated way and to guarantee a certain flexibility in planning to avoid under- and overutilization, e.g., in the form of cancellations. In this work, we introduce a performance-flexible artificial intelligence (AI)-based planning approach for predicting whether an elective patient will be transferred to the ICU after elective surgery. This approach includes a performance-flexible loss function in a machine learning (ML) model and a subsequent simulation about ICU occupancy. The algorithm is evaluated by a large data set of the University Hospital of Augsburg, Germany, consisting of more than 26,600 elective surgeries between 2017 and 2021, and extensive simulation studies. The data contains values that are determined in each hospital during the planning of the surgery, which is why this is a generalizable approach. We find that our performance-flexible AI-based planning algorithm, unlike state-of-the-art ML algorithms, can flexibly prioritize a particular label in binary classification (i.e., ICU or non-ICU) subject to capacity considerations while maintaining high accuracy. Consequently, the ratio of realized and planned intensive care resources is stable and close to 1 for different simulation scenarios regarding scarcity of resources and patient cohorts. Our performance-flexible AI-based planning algorithm outperforms state-of-the-art ML algorithms and supports hospital decision makers with a flexible planning tool.

Keywords: medical decision making, machine learning, binary classification, integrated capacity planning, loss function

1 Introduction

Operating room (OR) and intensive care unit (ICU) capacities belong to the scarcest resources in hospitals and strongly depend on each other: A considerable number of patients is transferred to the ICU after elective surgery (approx. 10% with sometimes high fluctuations). When planning elective surgeries, it is therefore important to consider both capacities in an integrated way, while allowing for some planning flexibility to accommodate scarcity patterns in both units (van Oostrum et al. 2008). Scarcity may be caused by staff shortages, mass casualties, or pandemics, for example (Heimerl and Kolisch 2010; Rodríguez-Espíndola 2023).

In recent years, and significantly pushed by the COVID-19 pandemic, planning and decision making in healthcare is on its way to be revolutionized by artificial intelligence (AI) approaches and in particular machine learning (ML) algorithms. ML-based predictions may also support the decision whether a patient is transferred to the ICU (i.e., ICU or non-ICU) after elective surgery. This decision is influenced by different stakeholders in the pathway of elective surgery patients (see Figure 1). About x days ($t = n - x$, where $20 \leq x \leq 30$) before elective surgery where n is the planned date of surgery, a patient is assigned an appointment in the surgical outpatient clinic. At this point in time, the surgeon provides a first assessment (D^2) whether the patient will need ICU treatment after surgery (i.e., ICU or non-ICU). The assessment might influence planning of the patient's surgery date. About 1 day ($t = n - 1$) before elective surgery, additional preoperative evaluation by anaesthesiologists is done. In this perioperative risk evaluation, the anesthetist also provides an assessment (D^1) on the need of ICU treatment after surgery (i.e., ICU or non-ICU). Additional findings and assessments by the anesthesiologist feed into the decision-making process by the surgeon and quite often change the decision with corresponding risk of cancellations. During the surgery ($t = n$), complications may also lead to a patient being transferred to the ICU although no ICU capacity was reserved for this patient (D^0). The pathway of elective surgery patients illustrates the potential of applying a ML model in each of the

three decision periods D^2, D^1 and D^0 for the prediction of actual ICU treatment after elective surgery to avoid cancellations. According to Heider et al. (2022), up to 35% of elective surgeries may be cancelled due to capacity issues in the ICU, which represents a fairly high potential of improvement. In the healthcare sector, a recurring issue is the higher initial demand for OR capacity compared to the available capacity. This prompts a proactive approach of aiming for complete utilization of this capacity beforehand. The subsequent focus shifts towards effectively managing the outcome of achieving 100% utilization, encompassing an integrated consideration of both the OR and ICU. Therefore, the concept of capacity is delineated as potential number of patients within the OR and the ICU. When the feasible patient number falls below the typical capacity in either OR or ICU, we describe it as scarce capacity. Conversely, if the potential patient number surpasses the norm, it signifies a state of high capacity. In our case, it is the scarcest situation when the OR capacity is high, and the ICU capacity is low. This can lead to the need for elective surgeries to be postponed.

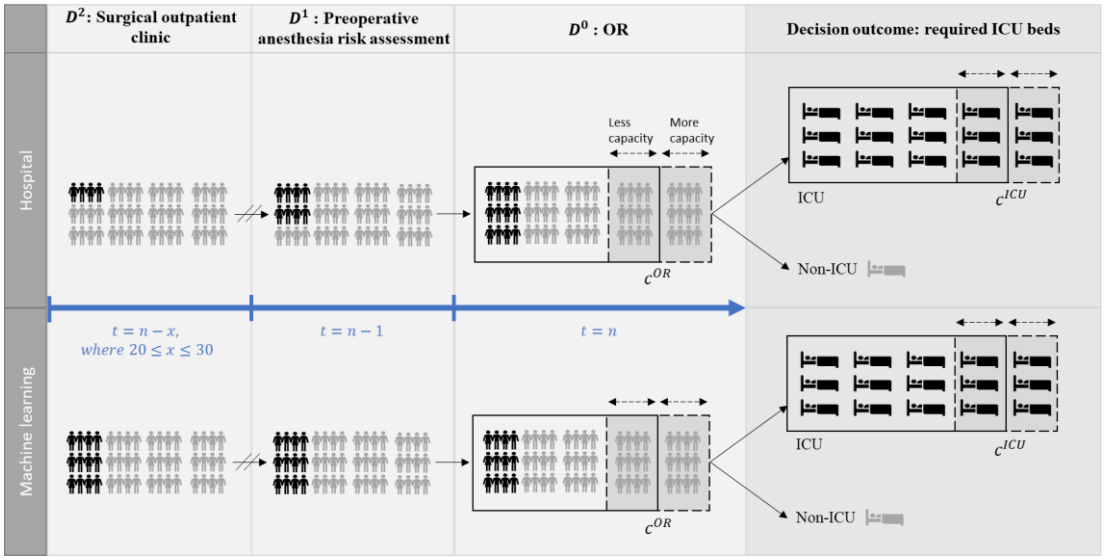

Figure 1: The pathway of elective surgery patients and different scarcity patterns in the OR and on ICU in the hospital (top) vs.ML algorithm (bottom). While in the current process in the hospital it is only clear after all three decision periods how many elective patients will be admitted to the ICU, ML can predict those already in the surgical outpatient clinic. This is particularly important when ICU beds are a scarce resource. For example, if there is little free capacity available in the ICU, it is even more important to achieve an accurate prediction, i.e., to achieve high sensitivity.

While state-of-the-art applications of ML methodologies concentrate on the accuracy of the overall prediction, the idea of integrating scarce OR and ICU capacities supposes planning flexibility and the dynamic focus, e.g., high sensitivity or high specificity, on a specific label such as ICU. If, for example, ICU capacity c^{ICU} is scarce while OR capacity c^{OR} is sufficient, it is particularly important to provide highly sensitive predictions while maintaining high specificity and accuracy. If ICU capacity is sufficient, while that of the OR is scarce, no particular focus on sensitivity is needed, which could increase the importance of specificity (see Figure 1).

In this work, we introduce a performance-flexible AI-based planning approach for predicting whether an elective patient will be transferred to the ICU after elective surgery (i.e., ICU or non-ICU). This approach includes a performance-flexible loss function in a ML model and a subsequent simulation about ICU occupancy. In the loss function, two weights are set for the respective focus on sensitivity or specificity. The approach is evaluated by a large data set of the University Hospital of Augsburg, Germany, consisting of more than 26,600 elective surgeries between 2017 and 2021 and extensive simulation studies. The data contains values that are determined in each hospital during the planning of the surgery, which is why this is a generalizable approach. We find that our performance-flexible AI-based planning approach, other than state-of-the-art ML algorithms, can flexibly prioritize a particular label in binary classification (i.e., ICU or non-ICU) subject to capacity considerations while maintaining high accuracy. Consequently, the ratio of realized demand and planned intensive care capacity is stable and near to 1 for different simulation scenarios regarding scarcity of resources and patient cohorts. Our performance-flexible AI-based planning approach supports decision makers in hospitals while guaranteeing planning flexibility.

Our work is structured as follows. In section 2, we provide an overview of related literature. In section 3, we introduce our performance-flexible AI-based planning approach. In particular, we describe the considered loss functions, thresholds, ML models, the data set and our simulation study. An

overview of the findings is provided in section 4. Section 5 includes a summary and an outlook to future research.

2 Related literature

ML has gained significant importance in recent years (e.g., Ratku and Neumann (2022)), especially in healthcare applications (e.g., Reig et al. (2020), Sheng et al. (2022)), and the number of publications using this methodology has increased accordingly. In this paper, we narrow our focus to binary classification, because we are interested in the binary decision of ICU treatment after elective surgery (i.e., ICU or non-ICU). We investigate two popular ML algorithms, namely Logistic Regression (LR) and Deep Neural Networks (DNN), to address this task. When applying ML methods to binary classification, various objectives can be pursued. These can be achieved by adjusting the components of the model. Well known components to be modified are the loss function and the threshold between the two classes. Existing literature introduces on the one hand new loss functions with the goal of data balancing and on the other hand problem-specific thresholds with different goals, like data balancing or priority on a performance measure (e.g., sensitivity). Essentially, the loss function is relevant during and after training, while the threshold is used after training, and thus after learning.

Cost-sensitive (learning). The objectives of data balancing or emphasizing different performance measures are often referred to as cost-sensitive learning. In this context, costs are not necessarily monetary, but can also represent the severity of a disease, for example. Under these circumstances, it might be more beneficial to predict the positive class, even if the negative class has a higher probability. The term cost-sensitive decision making is also used in this context (Elkan 2001). Cost-sensitive learning can be categorized among others into relabeling, weighting, and thresholding. In relabeling, classes of instances can be renamed based on cost. Weighting uses loss functions to assign a certain weight to each instance while thresholds are chosen to minimize the cost (Sheng and Ling 2006). In the following, we will consider

the cost-sensitive methods of weighting and thresholding in more detail. The notation for the following introduction of weighting can be seen in Table 1. An overview of all methods is given in Table 2.

Mathematical formulations	
N	Number of features
M	Number of training samples
P	Number of training samples of the positive class
y_m	Target label for training sample m
$\boldsymbol{x}_m$	Features for training sample m
$h_{\boldsymbol{\theta}}$	Prediction model with weights $\boldsymbol{\theta}$
α	Weight of positive label
β	Weight of negative label
γ	Learning rate

Table 1: Notation of loss functions

Weighting (Loss functions). An impactful and effective approach to influence ML models is the strategic manipulation of the loss function. Loss functions aim to minimize the error between label and prediction.

For the introduction and comparison of the different loss functions, we introduce a generic loss function for binary classification. Essentially, the loss function for two classes can be partitioned into two components. The first segment pertains to accurately predicting the positive class (true positives (TP)), whereas the second segment concerns precise predictions of the negative class (true negatives (TN)). The primary goal of the loss function is to reduce instances of false negatives (FN) in the first segment and false positives (FP) in the second segment. Moreover, it is possible to introduce weights to either or both segments. In our generic loss function, these weights are denoted as weight α for the positive class and weight β for the negative class:

$$J = -\frac{1}{M}\sum_{m=1}^{M}\left[\alpha \cdot y_m \cdot \log\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_m)\right) + \beta \cdot (1 - y_m) \cdot \log\left(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_m)\right)\right] \qquad (1)$$

The values of these weights change in the different loss functions, which will be discussed in more detail below. A well-known loss function in the area of binary classification is the *binary cross-entropy* (*BCE*) (Jadon 2020):

$$BCE: \alpha = 1, \beta = 1 \tag{2}$$

In this case, misclassification of the positive and negative class is penalized equally. Besides the state-of-the art approach of equal weights, some other loss functions based on *BCE* with a focus on binary classification already exist. The objective of these functions is to balance data.

Unbalanced data is a major problem in healthcare. For example, there is a greater abundance of data available from the normal ward as opposed to the ICU, or when predicting traumatic events compared to diseases with high prevalence rates. Therefore, many different methods of under- and oversampling have been developed over time to address this problem. An overview of existing methods is given by He and Garcia (2009). Among others, the authors mention the well-known Synthetic Minority Oversampling Technique (SMOTE) method to balance the data (Chawla et al. 2002). However, it should be noted that there is a debate on which method is superior (Drummond and Holte 2003). Another method of dealing with unbalanced data is to use modified loss functions. The *weighted binary cross-entropy (WBCE)* is widely used for unbalanced data. In binary classification, the false negative (FN) case is reinforced with a weight (Jadon 2020):

$$WBCE: \alpha \in \mathbb{R}_0^+, \beta = 1 \tag{3}$$

The weight for *WBCE* can be any positive real number. A higher weight indicates a higher prioritization on the correct prediction of the respective class. Studies show that the *WBCE* loss function causes a positive impact on model performance with unbalanced data (e.g., Rezeai-Dastjerehei et al. (2020)). In combination with LR, *WBCE* forms a special form called *weighted Logistic Regression*. This also aims to pay more attention to the minority class. There are many use cases (e.g., Das et al. (2013); Maalouf and Siddiqi (2014))

and even some publications in healthcare (e.g., Sheng et al. (2022); Zare et al. (2013)).

Unlike the *WBCE* loss function, *cost-sensitive binary cross-entropy* (*CSBCE)* uses solely the weight of the positive class and its convex weighting (Aurelio et al. 2019):

$$CSBCE: \alpha \in [0,1], \beta = (1 - \alpha) \tag{4}$$

The value of the weight of *CSBCE* can be between zero and one. Since the objective of the loss function, as motivated above, involves balancing the data, the weight can be determined by $\alpha = \left(\frac{P}{M}\right)^{-1}$. Thus, the ratio of the class and the total data is included in the loss function. A similar approach is taken by Cui et al. (2019) with their class-balanced loss function. There is no unique name for this loss function. Besides the *CSBCE*, for example, the name *balanced cross-entropy* is also a well-known term (Jadon 2020). In addition, there is the *focal loss* to balance the data. The weight in the loss function is determined by the class and the classification difficulty, resulting in lower weights assigned to examples that are easy to classify (Lin et al. 2017).

Furthermore, the weights may reflect the true cost of misclassification. This case is used in the *real-world-weight cross-entropy* (*RWWCE)* loss function. Among others, *RWWCE* can be applied to binary classification. The loss introduces one weight for the cost of missing a positive label, i.e., a FN, and a separate weight for missing a negative label, i.e., a false positive (FP). For binary classification, the weights of *RWWCE* are denoted as follows (Ho and Wookey 2020):

$$RWWCE: \alpha \in \mathbb{R}_0^+, \beta \in \mathbb{R}_0^+ \tag{5}$$

The weights can be any positive real number. All loss functions in literature target balancing data and balanced data plays an important role in healthcare.

Thresholding. In addition to influencing learning through the loss function, the outcome (i.e., the performance measurements) can also be changed through thresholding. The threshold parameter p allows for the conversion of predicted probabilities into specific labels based on a predetermined value. For binary classification, the state-of-the-art value is $h = 0.5$, which means that the positive class is predicted if the probability is greater than 0.5, and the negative class is predicted if the probability is smaller than or equal to 0.5:

$$predicted\ label = \begin{cases} positive\ label, if\ predicted\ probability > 0.5 \\ negative\ label, if\ predicted\ probability \leq 0.5 \end{cases} \quad (6)$$

If necessary, this threshold can be adjusted. For example, to focus on sensitivity, the threshold can be set to $h = 0.2$, since all probabilities above 0.2 will lead to a positive class prediction. This difference between setting different thresholds is shown in Figure 2. The figure visualizes 500 patients with two exemplary features, where the range of values can be neglected for understanding thresholds. The blue colors in the background show the different predicted probabilities. Darker colors indicate a higher predicted probability and vice versa.

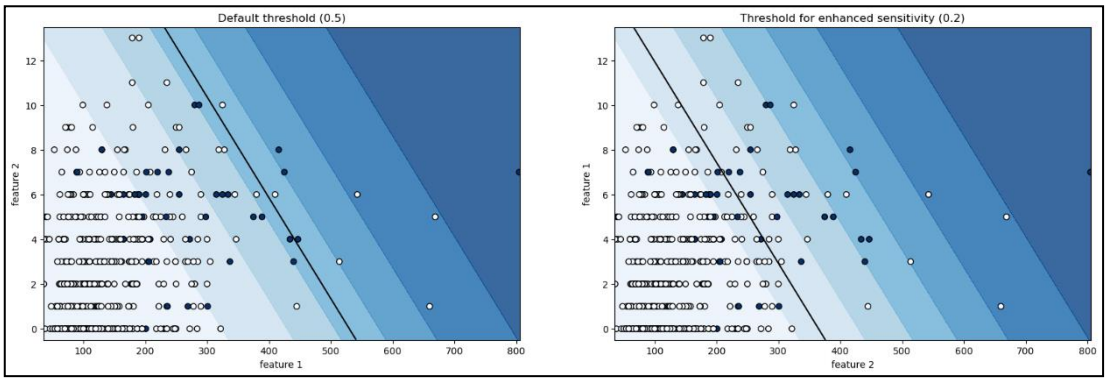

Figure 2: Default threshold and increased sensitivity threshold for LR and two features (left side of the line: negative label, right side of the line: positive label)

The white dots represent samples of the positive class, and the black dots represent samples of the negative class. The black line indicates the threshold at which the sample is classified as positive (right side of the line, e.g., ICU) or negative (left side of the line, e.g., non-ICU) based on the predicted

probabilities of the ML model. All samples on the right-hand side of the line are assigned to the positive class and all samples on the left-hand side are assigned to the negative class. By using the default state-of-the-art threshold of $h = 0.5$, the line is placed precisely at the midpoint ($predicted\ probability = 0.5$). If a higher sensitivity is desired, the line is moved to the left. This results in more samples being assigned to the positive class.

Individual thresholds allow algorithms to be made cost-sensitive even after training has been completed (Sheng and Ling 2006). Various objectives can be found in literature: Thresholding can be used to solve the problem of unbalanced data. Zhou and Liu (2006) show that sampling and thresholding achieve good results for unbalanced data sets, especially for binary classification. In addition, thresholding can be used for general cost-sensitive learning. It has been shown that thresholding almost always produces the lowest misclassification cost (Zhou, 2006). Furthermore, thresholding is used to maximize certain performance measures (i.e., sensitivity). Therefore, thresholding can be used to maximize the F1 score (Lipton et al. 2014), and flexible performance measure prioritization of classification problems (Eban et al. 2017).

Abbreviation	**Term**	**M**	**Parameter**	**Input**	**Objective**
BCE	*Binary cross-entropy*	-	$\alpha, \beta = 1$	-	-
BCET	*Binary cross-entropy with threshold*	T (NL)	$\alpha, \beta \in [0,1]$	UD, BD	BD, FLP
CSBCE	*Cost-sensitive binary cross-entropy*	W (L)	$\alpha, \beta \in \mathbb{R}_0^+$	UD	BD
NPFBCE	***Normalized performance-flexible binary cross-entropy***	**W** (L)	$\alpha, \beta \in [0,1]$	**BD**	**FLP**
PFBCE	***Performance-flexible binary cross-entropy***	**W** (L)	$\alpha, \beta \in [0,1]$	**BD**	**FLP**
RWWCE	*Real-world-weight cross-entropy*	W (L)	$\alpha, \beta \in \mathbb{R}_0^+$	UD	BD
WBCE	*Weighted binary cross-entropy*	W (L)	$\alpha \in \mathbb{R}_0^+, \beta = 1$	UD	BD

Table 2: Overview of methods including abbreviations, terms, and objectives; M: method (including learning (L) vs. no learning (NL) in training, T: thresholding, W: Weighting, FLP: flexible performance, BD: balanced data, UD: unbalanced data

Besides the achieved objectives of thresholds, the ease of implementation and application can also be highlighted (Sheng and Ling 2006). A major limitation

of thresholding is the lack of implementation in the training phase of the model. Consequently, the model cannot learn existing preferences. For example, the importance of features cannot be influenced by thresholding.

The literature review shows that while there are loss functions that address data balancing, no loss function addresses the goal of flexible performance. This goal is indeed addressed by another method, thresholding. However, thresholding is applied after the training phase of the model only. Thus, to the best of our knowledge, there is currently no method that integrates the objective of flexible performance with loss functions, which would impact the training of the model. We intend to close this research gap for our healthcare application.

3 Methodology

In the following, the new performance-flexible AI-based planning approach is introduced. For this purpose, we first present loss functions for binary classification, where the *CSBCE* is the basis for our objective and three other loss functions, *BCE* in combination with thresholding (*BCET)*, *WBCE* and *RWWCE,* are used as comparison approaches. Subsequently, the ML models used, and the data set applied are presented. A simulation is then performed to evaluate the new approach.

When predicting patients requiring critical care or a particular disease (i.e., ICU and non-ICU), it is essential for a healthcare decision maker to prioritize a specific class. As the relevance of a performance measure intensifies for decision makers, we first examine the confusion matrix and the three performance measures accuracy, sensitivity, and specificity. The distribution of correctly and incorrectly predicted cases can be determined based on the confusion matrix in Figure 3.

While accuracy determines the correct predictions among all predictions, sensitivity (specificity) calculates the true positive (TP) (true negative (TN)) predictions among all positive (negative) cases.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7}$$

$$sensitivity = \frac{TP}{(TP + FN)} \tag{8}$$

$$specificity = \frac{TN}{(FP + TN)} \tag{9}$$

These performance measures can be influenced in different ways in training, e.g., by loss functions, and in testing, e.g., by thresholding (see section 2).

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 3: Confusion matrix

Performance-flexible loss function. In this paper, we consider a supervised learning problem for binary classification of patients. With the help of performance measures, the output of the model can be measured. Standard models aim to achieve the highest possible accuracy, but there are many reasons to consider different ML models depending on the situation, e.g., a focus on sensitivity for predictions of ICU patients. The notation for the following introduction of loss functions can be seen in Table 1.

The objective of our performance-flexible AI-based planning approach is not cost minimization or data balancing. This requires a data set that has been previously balanced to focus on flexible performance. Our new approach uses the *CSBCE* loss function, except that it has a different objective and thus the weight is determined differently. We define our loss function as the *performance-flexible binary cross-entropy (PFBCE)*. In this approach, the

flexible performance can be influenced by the weight α. To emphasize both classes equally and focus on accuracy, the weight can be set to 0.5, which can be used as state-of-the-art approach.

$$PFBCE: \alpha \in [0,1], \beta = (1 - \alpha) \tag{10}$$

If a higher sensitivity is important for the decision maker, e.g., due to a low ICU capacity, a greater value than 0.5 can be assigned to α (e.g., $\alpha = 0.8$). This setting automatically downgrades the priority of correctly predicting the negative class, i.e., focus on specificity. Conversely, when ICU capacity is normal, a focus on the second class, i.e., non-ICU, can be chosen by using a lower weight of α (e.g., $\alpha = 0.2$). When considering extreme values in the lower range ($\alpha = 0.0$) and upper range ($\alpha = 1.0$), the focus is mainly on one class. This does not mean that the algorithm does not learn at all, but that the calculation of an error in the zero-weight class is not penalized. In general, nevertheless, it is not recommended to apply extreme values, i.e., $\alpha = 0$ and $\alpha = 1$.

The efficacy of the *PFBCE* approach can initially be justified through the following mathematical expressions. Like the *BCE*, the *PFBCE* is also a convex function, because an optimization approach like gradient descent should minimize the losses. For the *BCE*, the parameter update of the gradient descent would be calculated as follows:

Gradient descent BCE:

$$\theta \coloneqq \theta - \gamma \cdot \frac{1}{M} \sum_{m=1}^{M} [(h_{\boldsymbol{\theta}}(\boldsymbol{x}_m) - y_m) \cdot \boldsymbol{x}_m] \tag{11}$$

For the *PFBCE*, the convexity of the function is preserved by the convex approach:

Gradient descent PFBCE:

$$\theta \coloneqq \theta - \gamma \cdot \frac{1}{M} \sum_{m=1}^{M} [(\alpha \cdot h_{\boldsymbol{\theta}}(\boldsymbol{x}_m) - (1-\alpha) \cdot y_m) \cdot \boldsymbol{x}_m] \tag{12}$$

Normalization. Due to anomalies in the data (e.g., different features, biased values) and a lack of comparability of weighted loss functions, normalizing the loss function can be useful from a practical perspective. First, normalization can prevent large differences between different features in a data set and avoid problems with biased values (Ma et al. 2020). Second the problem of comparability can be avoided by applying normalization. In this work, normalization always refers to the normalization of the loss function. One way to do this is to divide the *PFBCE* by the sum of the weights for each class. For this purpose, we introduce the *normalized performance-flexible binary cross-entropy (NPFBCE)* loss function:

$$J_{NPFBCE} = -\frac{1}{M} \sum_{m=1}^{M} \frac{\left[\alpha \cdot y_m \cdot log\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_m)\right) + \beta \cdot (1-y_m) \cdot log\left(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}_m)\right)\right]}{\alpha \cdot y_m + \beta \cdot (1-y_m)} \tag{13}$$

The user of the ML algorithm can decide whether to normalize the loss function by changing the corresponding hyperparameter. This grants the flexibility to adapt the loss function based on the situation and, consequently, the data set.

Machine learning models. Since loss functions are used in several supervised learning methods, we will use LR and a DNN as examples. We decided to use these two models because LR is often used together with weighted LR, and DNN is probably the most popular ML model performing well for many applications (Cichy and Kaiser 2019). For sake of comparability, the same ML models are applied to all loss functions. For the DNN architecture, hyperparameter tuning was first performed for the *NPFBCE* loss function with $\alpha = 0.8$. The weight of 0.8 was used because our application

requires a high focus on sensitivity. This results in a DNN with one input layer, six hidden layers, and one output layer. Except for the output layer, where the sigmoid activation function was applied, the tanh activation function is used. The exact DNN is shown in Figure 4.

For training, we use the loss functions *BCET*, *NPFBCE*, *PFBCE*, *RWWCE*, *WBCE* presented above and stochastic gradient descent as the optimizer. Additionally, we use 85 epochs and a batch size of 10. To avoid overfitting, we use 5-fold cross validation. For a detailed evaluation of the results, we use the metrics accuracy, sensitivity, and specificity.

For the comparison of the loss functions some assumptions had to be made. The main point to be noted is that the two loss functions, *WBCE* and *RWWCE*, do in fact include the objective of balanced data. In this case, the pre-balanced data set leads to a distinct initial state and, as a result, a different objective. Moreover, *(N)PFBCE* can be understood as a subfunction of *RWWCE*, since this loss function can be set to the same values (e.g., $\alpha = 0.6$ and $\beta = 0.4$). However, this is not given in the context that the weight represents the estimated impact in the real world, and it is normally a real number.

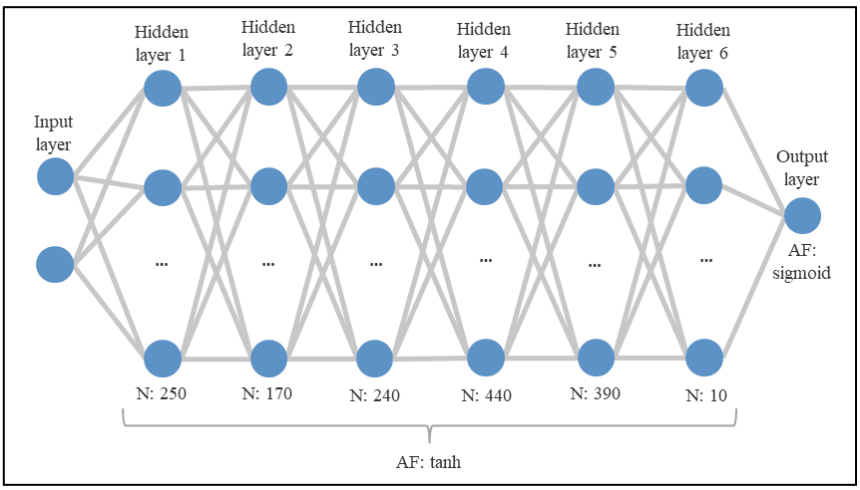

Figure 4: Overview of the DNN applied to the data set (AF: activation function, N: neurons)

Data and simulation methodology. Both LR and DNN are applied to a real-world healthcare data set. This data (2017 to 2021) contains 26,677 patients from the University Hospital of Augsburg, Germany, for whom a

decision must be made whether they require intensive care after elective surgery. All data is available at the first decision period D^2 in the surgical outpatient clinic. In the data set, approx. 10% of elective patients require subsequent ICU treatment. The data set includes 14 features, which are shown in Table 3.

No.	Feature	No.	Feature	No.	Feature
1	Medical specialty	6	Sex	11	Number of comorbidities
2	Estimated cut suture duration	7	Weight	12	CCI
3	Estimated anesthesia duration	8	Height	13	Planned type of anesthesia
4	Estimated surgery duration	9	Body mass index	14	Estimated ASA-score
5	Age	10	Main diagnosis		

Table 3: Description of the 14 features of the data set (CCI: Charlson Comorbity Index)

Every planned surgery includes the binary label of whether a patient is admitted to the ICU afterwards. As with most healthcare data sets, extensive data preparation was required. Non-relevant or erroneous data were deleted, and comorbidities were summarized in the Charlson Comorbidity Index (CCI) (Charlson et al. 1987). We use the Synthetic Minority Oversampling Technique (SMOTE) to balance the data (Chawla et al. 2002). In addition, we use feature scaling and missing values are imputed with the Iterative Imputer, predicting missing values using the Random Forrest algorithm. The split of training and test data is 90% to 10%.

The major goal of the simulation study is to evaluate varying weights in the performance-flexible AI-based planning approach for different scenarios. Our 9 scenarios differ in scarcity patterns in the OR and ICU (see Table 4). The scarcity patterns have been estimated based on the data set introduced above of the central ORs at the University Hospital of Augsburg, Germany.

Scenario	OR capacity [surgeries]		
Ratio of ICU patients	20 (scarce)	30 (normal)	40 (high)
0.05 (scarce)	1	4	7
0.10 (normal)	2	5	8
0.20 (high)	3	6	9

Table 4: Scenarios of the simulation study

Scarce/normal/high OR capacity is defined to be 20/30/40 elective surgeries, i.e., elective patients, per day; scarce/normal/high ICU capacity is defined to be a ratio of 5/10/20 percent ICU patients per day. In addition, we defined 3 different patient cohorts: All patients and patients with ASA-scores[1] smaller than 3 or bigger than 2. The different patient groups cover different status of the healthcare system. For example, during the COVID-19 pandemic, authorities banned postponable surgeries of non-severely ill patients, who are simultaneously assumed to have low ASA-scores in this work. For every weight, scarcity pattern and patient group, Monte-Carlo simulations with 1,000 runs are applied. Per simulation run, we simulated 2 types of patients with replacement out of the given data: Patients that are predicted to be transferred to the ICU after elective surgery (i.e., ICU label) and patients that are predicted to not be transferred to the ICU after elective surgery (i.e., non-ICU). We calculated the ratio of ex-post realized and planned ICU patients as our key performance indicator (KPI) per simulation run and averaged the results for every weight, scarcity pattern and patient group. If the KPI is 1, realized ICU occupancy is the same as predicted. Consequently, a ratio of 1 is our benchmark. If the KPI is bigger than 1, less ICU occupancy than needed is planned. If the KPI is smaller than 1, more ICU occupancy than needed is planned. Please find a flowchart of the simulation study in Figure 5.

[1]ASA-score is a extensively employed scoring system utilized for categorizing patients based on their physical condition (Saklad (1941)).

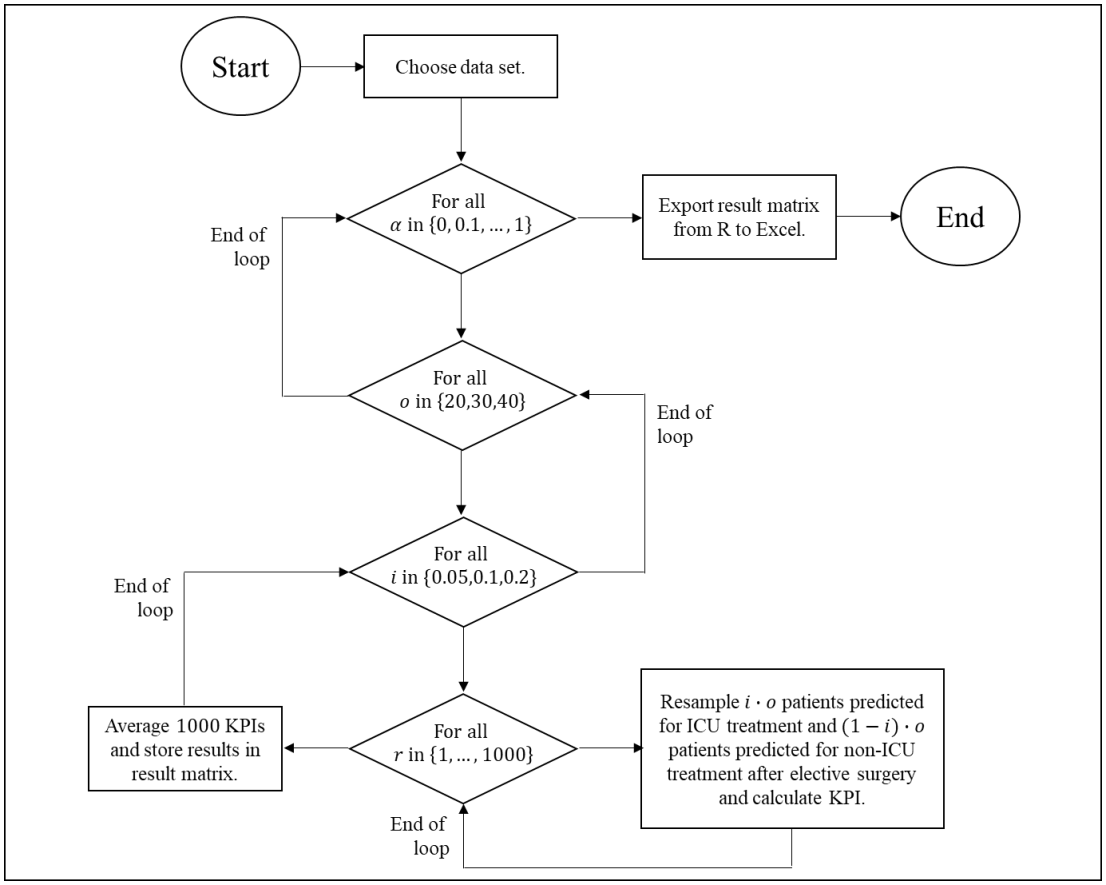

Figure 5: Flowchart of the simulation study; α: weight of positive class, o: OR capacity, i: ICU capacity, r: index for runs

For the performance-flexible AI-based approach, we used Python for the ML models and R for the simulation.

4 Results

In the following, we present our results by applying the methods to the ICU and elective surgery data. First, the results of the two ML models, LR and DNN are discussed. Afterwards, we present the simulation results with a focus on ICU capacity management. Below, absolute values are expressed as percentages (%) and differences are presented as percentage points (*PP*).

4.1 Comparing LR with DNN

In addition to our two loss functions *PFBCE* and *NPFBCE*, the methods *BCET*, *WBCE* and *RWWCE* presented in the literature section are used for

comparison. Our application of the *NPFBCE* loss function to the data set confirms the assumptions made above. By using a higher weight, a higher sensitivity can be achieved, which decreases the specificity. Conversely, increased specificity directly causes decreased sensitivity. However, there is no significant change in accuracy. The results of the LR and DNN are also shown in Figure 6.

On the one hand, this means that a higher weighting of *NPFBCE* generally allows a focus on one performance measure, namely sensitivity or specificity. On the other hand, this can assist a healthcare decision maker in capacity planning. For example, if ICU capacity is scarce, few ICU patients may be admitted after elective surgery. Therefore, in such a situation, it is particularly important to achieve high accuracy of the positive class, i.e., sensitivity. This is possible with the use of ML and adjustments of weights of the *NPFBCE*. If capacity is scarce, the weight can be set high (e.g., $\alpha = 0.8$). With a weight of 0.8, *NPFCE* in the DNN (in LR) achieves a sensitivity of 98.29 % (97.73 %), a specificity of 76.78 % (78.31 %), and an accuracy of 87.53 % (88.02 %).

The example shows that in both models a strong focus on one performance measure, e.g., sensitivity, still guarantees high accuracy. If the capacity is high, there is no need to focus on sensitivity. For example, the base case with a weight of 0.5, corresponding to *BCE* and state-of-the-art ML approaches, can be used. This results in a sensitivity of 93.08 % (91.33 %), a specificity of 85.46 % (86.65 %) and an accuracy of 89.27 % (88.99 %) for *PFBCE* for the DNN (LR). The use of the extreme values, i.e., $\alpha = 0.0$ or $\alpha = 1.0$, is not recommended because either the sensitivity or the specificity is zero. As the results of the LR are similar to those of the DNN, but especially the sensitivity of the latter is higher, we focus on the DNN for further analyses.

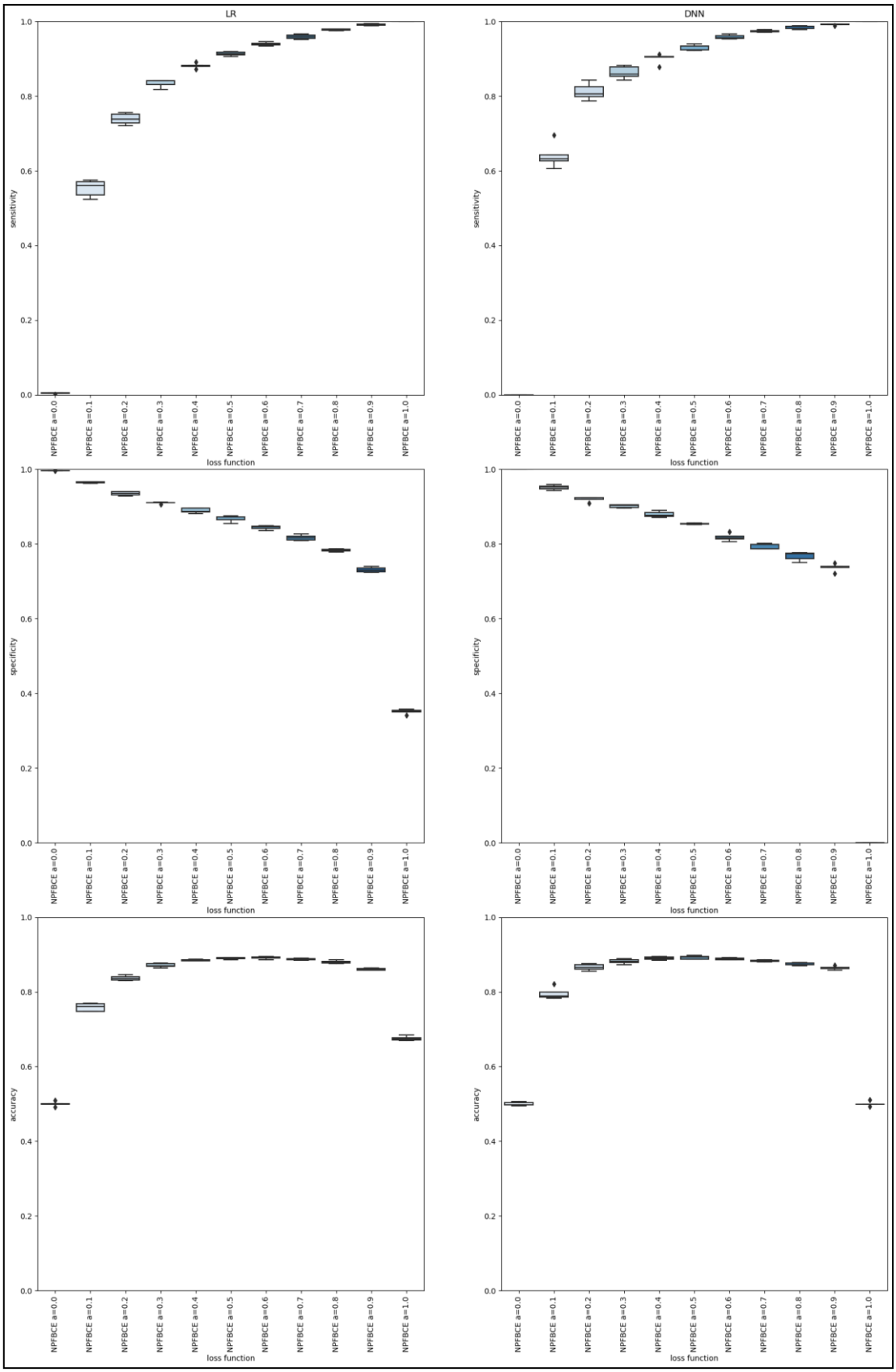

Figure 6: Performance measures sensitivity, specificity, and accuracy of LR and DNN for NPFBCE

Until now, we have solely examined the results of the *NPFBCE* loss function. However, these findings can also be applied to the non-normalized *PFBCE* loss function. When comparing the *PFBCE* loss function with the normalized version, there is no distinct difference. For example, for $\alpha = 0.7$ and the DNN, we obtain a sensitivity of 97.37 % (95.92 %), a specificity of 79.24% (81.68 %) and an accuracy of 88.31 % (88.80 %) for *NPFBCE* (*PFBCE*). In the remainder, we use the results of *NPFBCE* for the simulation performed in the following section.

A comparison of the results of the (*N)PFBCE* loss functions with the loss functions in literature confirms our previous findings. We consider, for example, high sensitivity, i.e., $\alpha = 0.8$. In DNN, *NPFBCE* (*PFBCE*) achieves an improved sensitivity of $+7.05\,PP$ $(+7.04\,PP)$, a reduced specificity of $-10.32\,PP$ $(-9.70\,PP)$ and a reduced accuracy of $-1.65\,PP$ (-1.33) compared to *WBCE*. This trend is also evident when comparing *(N)PFBCE* with *RWWCE*. These results show that our approach places higher emphasis on sensitivity than the loss functions *WBCE* and *RWWCE*, while maintaining comparable accuracy. Only the *BCET* achieves similar results to the *(N)FPCE* function in terms of sensitivity. With weight $\alpha = 0.8$, *NFBCE* (*PFBCE*) shows a decreased sensitivity $-2.56\,PP$ $(-3.04\,PP)$, an increased specificity $+5.91\,PP$ $(+6.80\,PP)$ and an increased accuracy $+1.68\,PP$ $(+1.88\,PP)$ compared to *BCET*. For sensitivity, *BCET* is slightly superior. The methodological disadvantage of *BCET*, compared to our performance-flexible AI-based planning approach, is that sensitivity in *BCET* is determined after training. Thus, learning of the model cannot be influenced by the given sensitivity. This aspect is important not only for learning patterns applicable to other data but also for the application of techniques such as feature importance, which is performed during training. The performance of the different loss functions with $\alpha = 0.6$ and $\alpha = 0.8$ is shown in Figure 7.

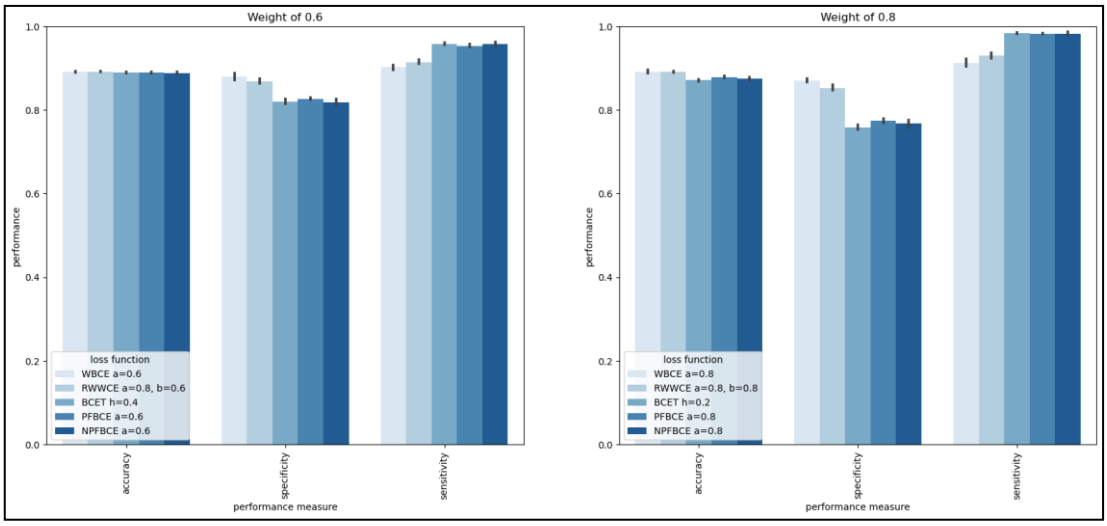

Figure 7: Comparison of loss functions for $\alpha = 0.6$ and $\alpha = 0.8$ of DNN

The results show that the *(N)PFBCE* loss function should be used in management decisions involving capacity constraints. Besides the focus on sensitivity, the function also outperforms the loss functions *WBCE* and *RWWCE*. *BCET* is quite similar in terms of performance, but has methodological weaknesses compared to the *(N)PFBCE* loss function. These findings can be transferred to the other weights with a focus on specificity, i.e., a weight smaller than 0.5. Additionally, with a high focus on sensitivity, *NPFBCE* performs better than *PFBCE* in our data set.

The performance measures of all loss functions for LR and DNN functions are shown in the Appendix. Due to memory constraints, only certain weights were considered for the two loss functions *WBCE* and *RWWCE*.

4.2 Evaluation of different capacity patterns using simulation

In the following, we discuss the outcomes of the simulation study for the different data sets, capacity patterns, i.e., simulation scenarios defined in Table 4, and weights. As discussed before, we focus on one KPI which is the ratio of ex-post realized and planned ICU patients. A KPI of 1 (benchmark) denotes that for all patients in need of ICU treatment after elective surgery sufficient treatment is available.

For the data set with all patients, normal ICU and OR capacity (scenario 5), we find a KPI of 1 for weight in between 0.8 and 0.9. A KPI of 1 is thus achieved only for a considerable focus on sensitivity in case of both normal ICU and OR capacity. For state-of-the-art ML with a weight of 0.5, the KPI is 1.29. This means that we expect 1.29 patients per unit of ICU capacity, and ICU capacity is overbooked. Consequently, based on the state-of-the-art ML approach, surgeries are to be postponed or cancelled or ICU patients must be discharged early which might endanger their health. In addition, the decision on the different options to meet overutilization causes significant additional managerial effort. This effort is avoided by applying our new approach with weights in between 0.8 and 0.9. The conclusions apply accordingly to the data set with all patients, low/normal/high ICU and low/high OR capacities. Consequently, performance-flexible AI-based planning of elective surgeries is important for ICU capacity usage and outperforms state-of-the-art ML predictions in all scenarios. Depending on the capacity patterns in the OR and in ICU, our new approach with weights in between 0.7 and 1 should be applied. Performance-flexible AI-based planning is thereby of special importance for scarce ICU capacities, while scarcity in the OR plays a subordinate role for performance-flexible ICU planning. For both scarce ICU and OR capacities, the gap between state-of-the-art planning and our performance-flexible AI-based planning approach becomes bigger but superior weights remain unchanged compared to normal OR.

When comparing these results for the data set with all patients with the data sets with high/low ASA-score patients, it becomes evident, that performance-flexible AI-based planning is of special importance for patient cohorts with average ASA-scores (see above) and high ASA-scores. For the data set with high ASA-score patients, weights in between 0.9 and 1 lead to a KPI near to 1. In contrast, if state-of-the-art planning is applied to this cohort, per unit of ICU capacity, we expect up to 2.80 patients. The data set with high ASA-score patients mimics, e.g., the situation during peak phases of the COVID-19 pandemic, when only severely ill patients had access to elective surgeries. If state-of-the-art planning is applied here, ICU capacities are significantly

overbooked with corresponding consequences on postponements, cancellations, early discharges, and management decisions. It becomes evident that the application of performance-flexible AI-based planning of elective surgeries is even more important in pandemic and similar circumstances.

For the data set with low ASA-score patients, in case of state-of-the-art planning, we expect only 1.10 patients per unit of ICU capacity. For the data set with low ASA-score patients and scenario 6 (4), i.e., high (scarce) ICU and normal OR capacity, the KPI is 1 for weights in between 0.6 and 0.7 (0.8 and 0.9). Consequently, state-of-the-art planning might be applied in the unlikely situation of rather healthy patients. As our performance-flexible approach can depict this unlikely situation and flexibly switch to other circumstances, the simulation study strongly supports the application of our integrated approach for ex-ante ICU planning of elective surgeries. A summary of the results of the simulation study is shown in Figure 8.

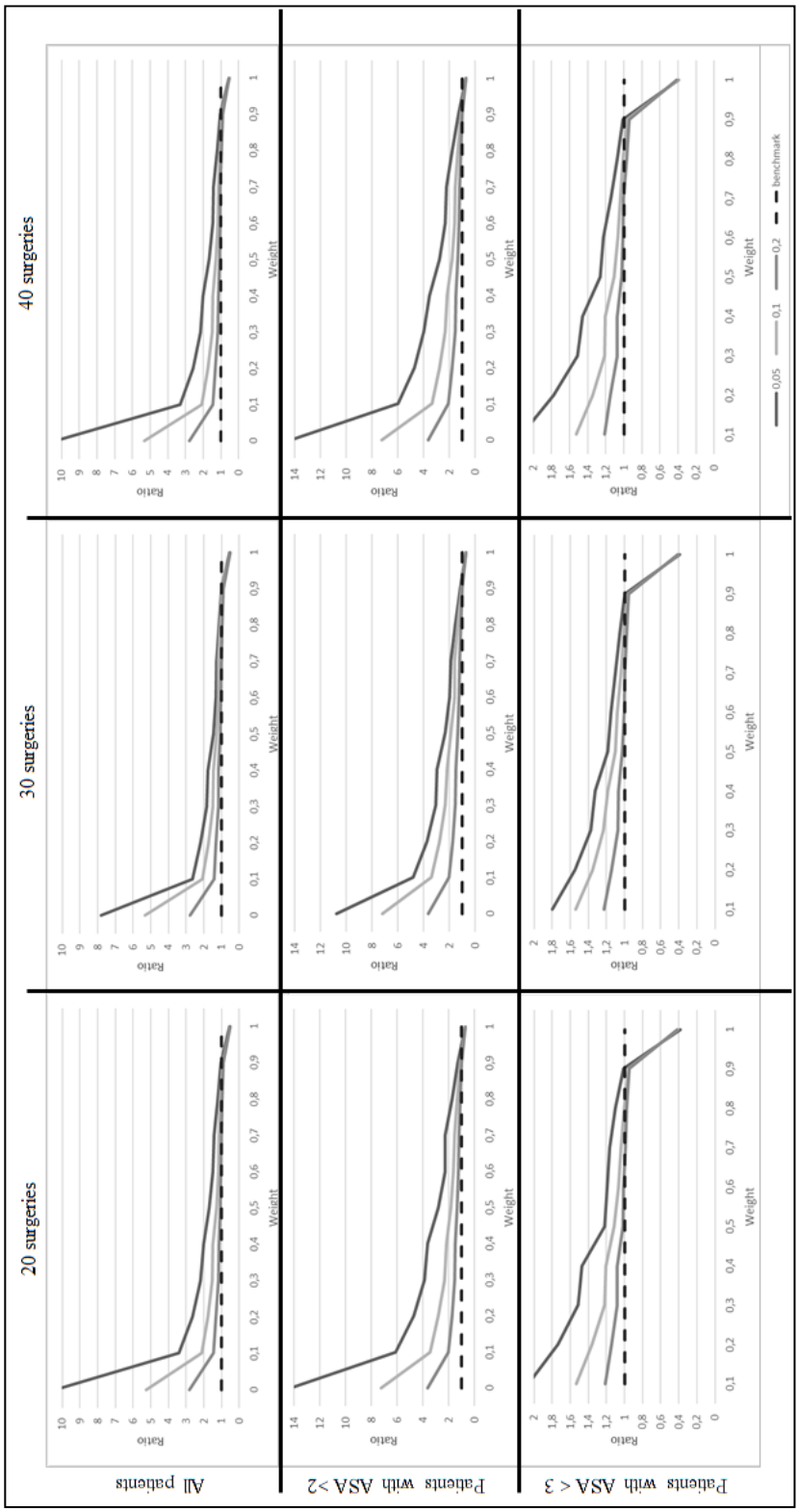

Figure 8: Ratios of ex-post realized and planned ICU patients (KPI) for the different simulation scenarios (see Table 4) and data sets.

5 Conclusion

This study introduces an innovative performance-flexible AI-based planning approach designed to predict ICU treatment following elective surgeries. Our approach emphasizes performance flexibility, achieved through a weighted loss function during training. By prioritizing a specific label while maintaining high accuracy, the model allows decision-makers, for example, to focus on sensitivity when ICU capacity is of crucial importance, leading to more accurate positive class predictions. The method considers both OR and ICU capacity for elective patients requiring post-surgery ICU care, resulting in stable ICU capacity ratios close to 1 across various simulation scenarios and patient cohorts with different resource availability.

There are some limitations to our methodology. First, data preparation steps applied to the data set can influence the model's results, and the impact of different preparation approaches remains unclear. Second, the loss functions for comparison include other objectives, leading to different requirements in implementation. Third, while the performance-flexible AI-based planning approach excels in the integrated ICU and OR case, its performance in other applications warrants further investigation.

Future research could extend our approach to address short-term scheduling. We focus on a planning horizon of approximately 30 days in advance, but short-term adjustments may be necessary. Additionally, applying the performance-flexible AI-based planning approach to other decision problems offers promising avenues for further exploration and application in diverse domains. Furthermore, the combination of the *(N)PFBCE* and thresholding could provide fruitful insights.

References

Aurelio YS, Almeida GM de, Castro CL de, Braga AP (2019) Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. Neural Process Lett 50:1937–1949. https://doi.org/10.1007/s11063-018-09977-1

Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Diseases 40:373–383. https://doi.org/10.1016/0021-9681(87)90171-8

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. jair 16:321–357. https://doi.org/10.1613/jair.953

Cichy RM, Kaiser D (2019) Deep Neural Networks as Scientific Models. Trends in Cognitive Sciences 23:305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cui Y, Jia M, Lin T-Y, Song Y, Belongie S (2019) Class-Balanced Loss Based on Effective Number of Samples. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition:9268–9277. https://doi.org/10.1109/cvpr.2019.00949

Das S, Moore T, Wong W-K, Stumpf S, Oberst I, McIntosh K, Burnett M (2013) End-user feature labeling: Supervised and semi-supervised approaches based on locally-weighted logistic regression. Artificial Intelligence 204:56–74. https://doi.org/10.1016/j.artint.2013.08.003

Drummond C, Holte RC (2003) Class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Workshop on Learning from Imbalanced Datasets II

Eban E, Schain M, Mackey A, Gordon A, Rifkin R, Elidan G (2017) Scalable Learning of Non-Decomposable Objectives. Artificial Intelligence and Statistics:832–840

He H, Garcia EA (2009) Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 21:1263–1284. https://doi.org/10.1109/tkde.2008.239

Heider S, Schoenfelder J, Koperna T, Brunner JO (2022) Balancing control and autonomy in master surgery scheduling: Benefits of ICU quotas for recovery units. Health Care Manag Sci 25:311–332. https://doi.org/10.1007/s10729-021-09588-8

Heimerl C, Kolisch R (2010) Scheduling and staffing multiple projects with a multi-skilled workforce. OR Spectrum 32:343–368. https://doi.org/10.1007/s00291-009-0169-4

Ho Y, Wookey S (2020) The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access 8:4806–4813. https://doi.org/10.1109/access.2019.2962617

Jadon S (2020) A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE

Lin T-Y, Goyal P, Girshick R, He K, Dollar P (2017) Focal Loss for Dense Object Detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE

Lipton ZC, Elkan C, Narayanaswamy B (2014) Thresholding Classifiers to Maximize F1 Score. arXiv preprint arXiv 1402

Ma X, Huang H, Wang Y, Romano S, Erfan S, Bailey J (2020) Normalized Loss Functions for Deep Learning with Noisy Labels. International Conference on Machine Learning:6543–6553

Maalouf M, Siddiqi M (2014) Weighted logistic regression for large-scale imbalanced and rare events data. Knowledge-Based Systems 59:142–148. https://doi.org/10.1016/j.knosys.2014.01.012

Ratku A, Neumann D (2022) Derivatives of feed-forward neural networks and their application in real-time market risk management. OR Spectrum 44:947–965. https://doi.org/10.1007/s00291-022-00672-1

Reig B, Heacock L, Geras KJ, Moy L (2020) Machine learning in breast MRI. Journal of Magnetic Resonance Imaging 52:998–1018. https://doi.org/10.1002/jmri.26852

Rezeai-Dastjerehei MR, Mijani A, Fatemizadeh E (eds) (2020) Adressing Imbalance in Mulit-Label Classification Using Weighted Cross Entropy Loss Function. IEEE

Rodríguez-Espíndola O (2023) Two-stage stochastic formulation for relief operations with multiple agencies in simultaneous disasters. OR Spectr 45:477–523. https://doi.org/10.1007/s00291-023-00705-3

Saklad M (1941) Grading of patients for surgical procedurs. Anesthesiology 2:281–284. https://doi.org/10.1097/00000542-194105000-00004

Sheng J, Wu S, Zhang Q, Li Z, Huang H (2022) A Binary Classification Study of Alzheimer's Disease Based on a Novel Subclass Weighted Logistic Regression Method. IEEE Access 10:68846–68856. https://doi.org/10.1109/access.2022.3186888

Sheng VS, Ling CX (2006) Thresholding for Making Classifiers Cost-sensitive. Aaai 6:476–481. https://doi.org/10.1016/B978-0-12-802021-0.00010-3

van Oostrum JM, van Houdenhoven M, Hurink JL, Hans EW, Wullink G, Kazemier G (2008) A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum 30:355–374. https://doi.org/10.1007/s00291-006-0068-x

Zare N, Haem E, Lankarani KB, Heydari ST, Barooti E (2013) Breast cancer risk factors in a defined population: weighted logistic regression approach for rare events. J Breast Cancer 16:214–219. https://doi.org/10.4048/jbc.2013.16.2.214

Zhou Z-H, Liu X-Y (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 18:63–77. https://doi.org/10.1109/TKDE.2006.17

Appendix

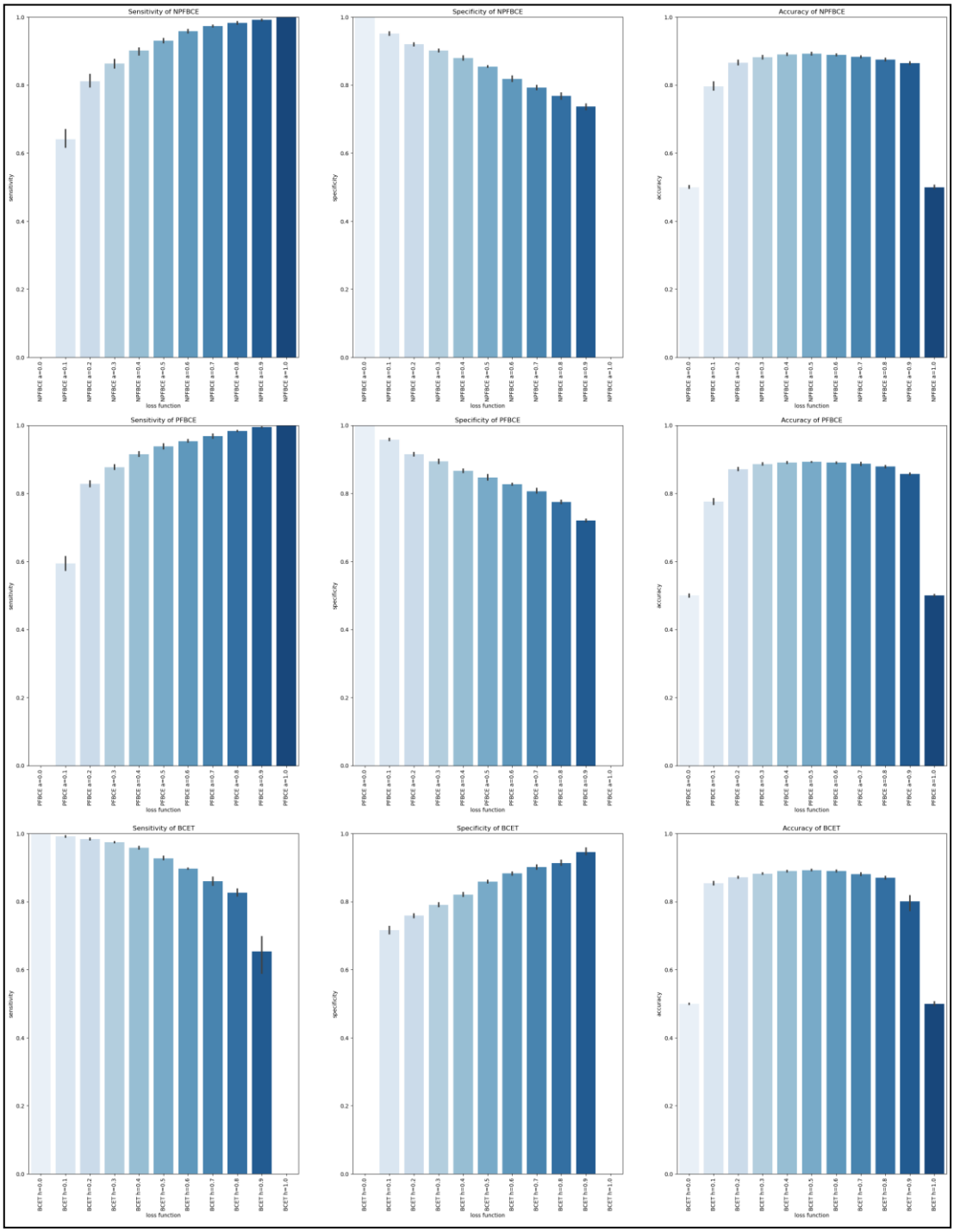

Figure 9: Performance measures sensitivity, specificity, and accuracy of NPFBCE, PFBCE and BCET of DNN

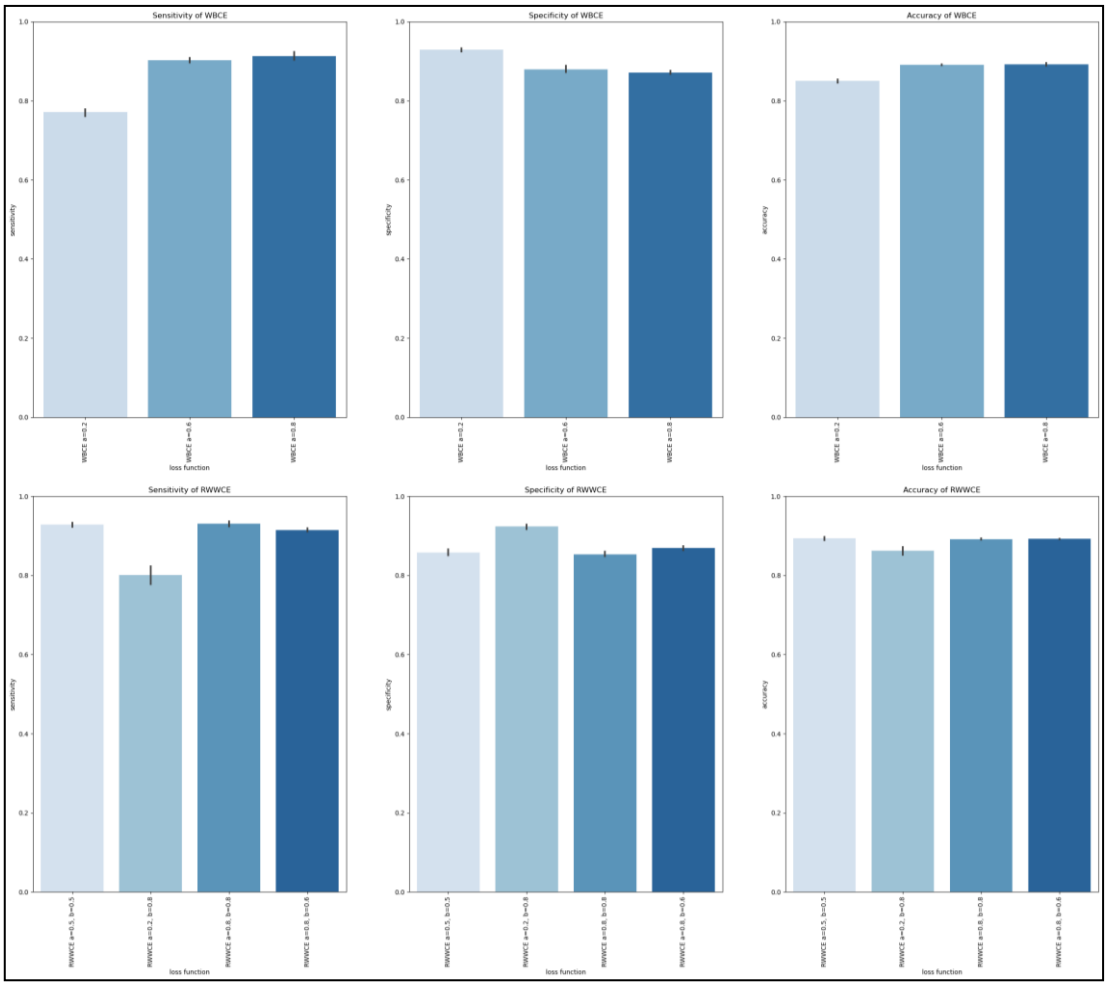

Figure 10: Performance measures sensitivity, specificity, and accuracy of WBCE and RWWCE of DNN

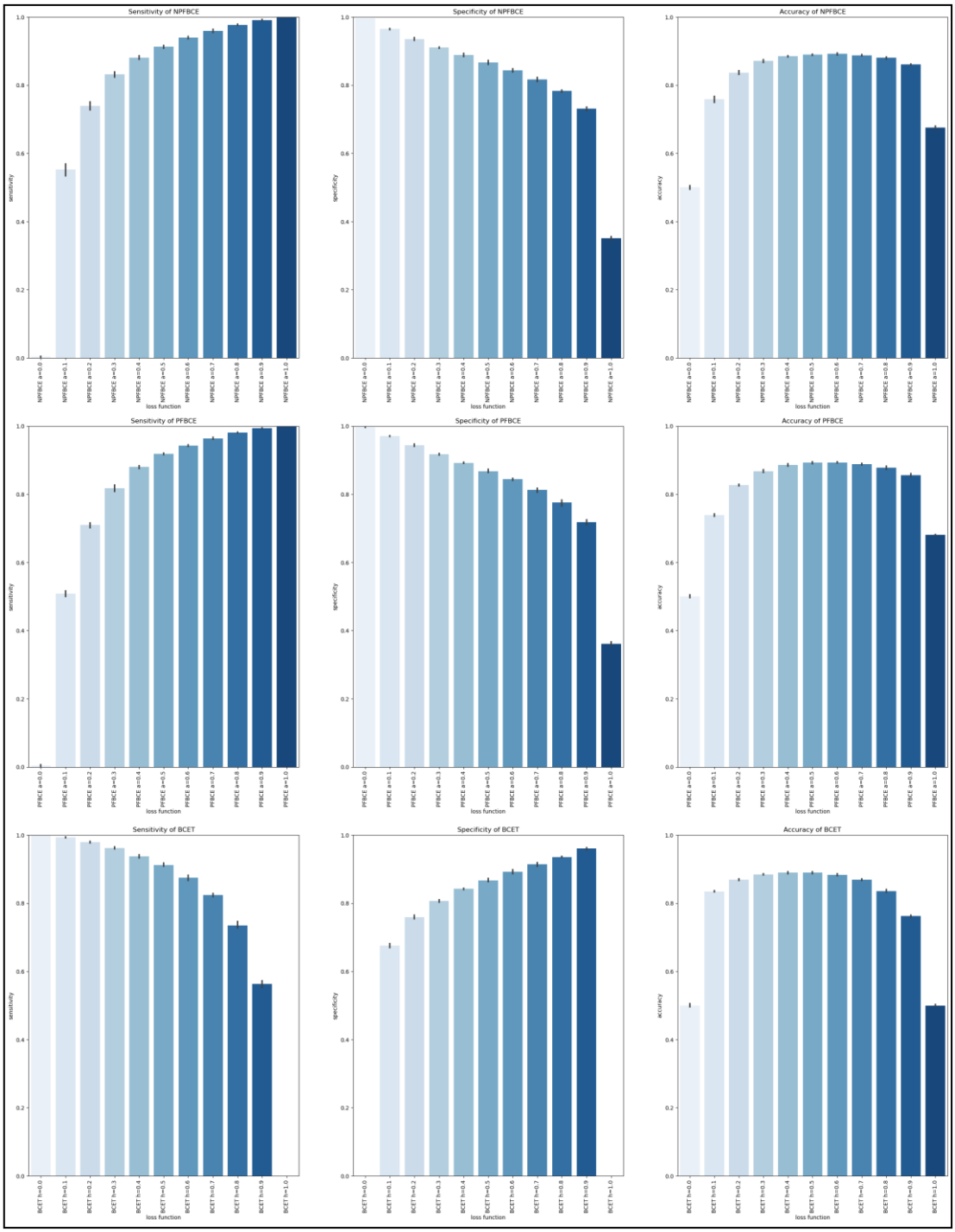

Figure 11: Performance measures sensitivity, specificity, and accuracy of NPFBCE, PFBCE and BCET of LR

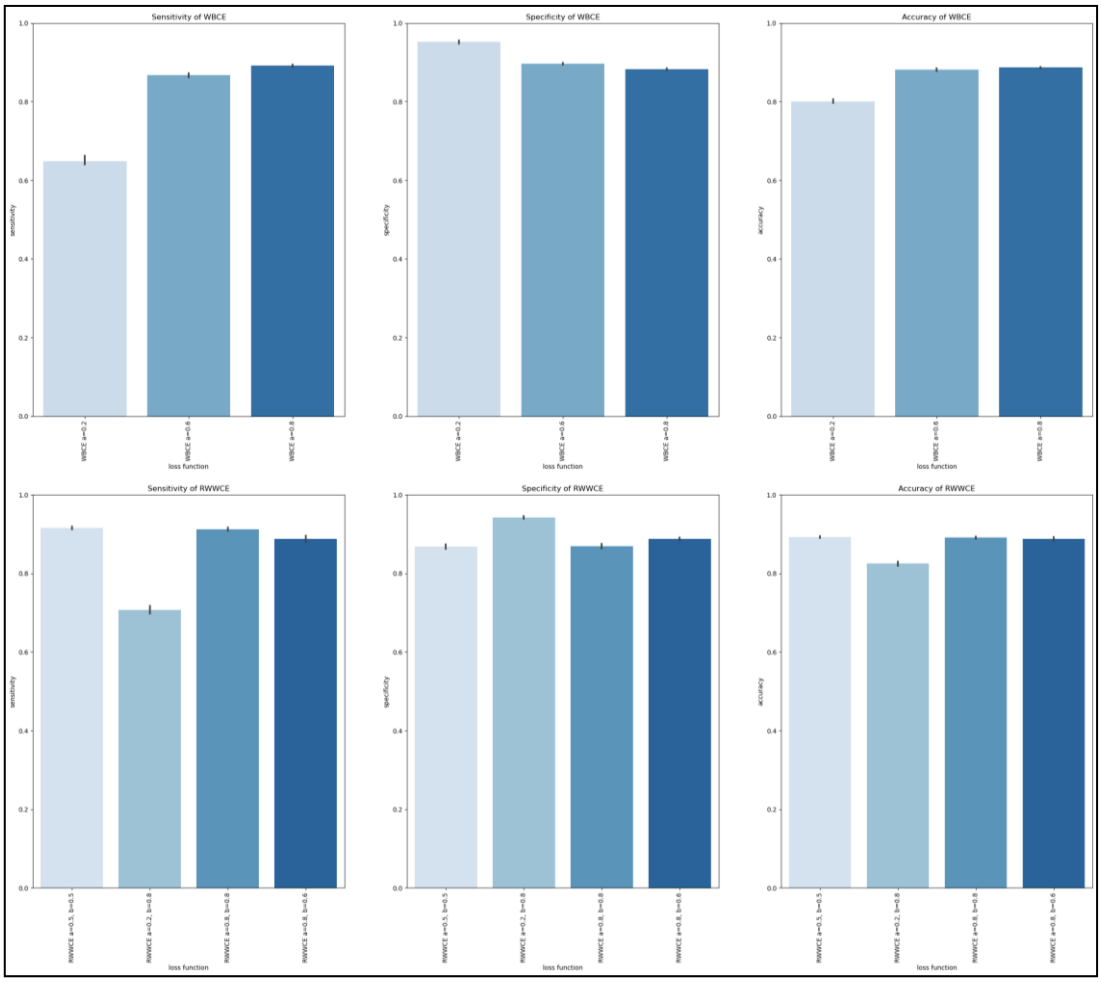

Figure 12: Performance measures sensitivity, specificity, and accuracy of WBCE and RWWCE of LR

Appendix B: Customized AFs

Grieger, M, Shala, E, Schüller, M, Ebel, SS, Brunner, JO, Vehreshild, JJ, Erber, J, Hanses, F, Zabel, LT, Römmele, C, Shmygalev, S, Bartenschlager, CC (2023). *DENLU* and *leaky stanh*: customized activation functions targeting enhanced sensitivity with healthcare applications in binary classification.

Status: Close to submission.

Original Research Paper:

DENLU and *leaky stanh*: customized activation functions targeting enhanced sensitivity with healthcare applications in binary classification

Milena Grieger[1], Elion Shala[1], Markus Schüller[1], Stefanie S. Ebel[1], Jens O. Brunner[1,10,11], Jörg J. Vehreschild[2,3,4], Johanna Erber[5], Frank Hanses[6], Lutz T. Zabel[7], Christoph Römmele[8], Sergey Shmygalev[9], Christina C. Bartenschlager[9,12]

[1]Health Care Operations/Health Information Management, Faculty of Business and Economics, Faculty of Medicine, University of Augsburg, Universitätsstraße 16, 86159 Augsburg, Germany
[2]Goethe University Frankfurt, Department of Internal Medicine, Hematology and Oncology, Frankfurt am Main, Germany
[3]University of Cologne, University Hospital of Cologne, Department I of Internal Medicine, Cologne, Germany
[4]German Center for Infection Research, partner site Bonn-Cologne, Cologne, Germany
[5]Technical University of Munich, School of Medicine, University Hospital rechts der Isar, Department of Internal Medicine II, Ismaninger Str. 22, 81675 Munich, Germany.
[6]Hygiene and Infectiology, University Hospital of Regensburg, Germany
[7]Laboratory Medicine, Alb Fils Kliniken GmbH, Eichertstraße 3, 73035 Göppingen
[8]Clinic for Internal Medicine III - Gastroenterology and Infectious Diseases, University Hospital Augsburg, Stenglinstraße 2, 86156 Augsburg, Germany
[9]Department of Anaesthesiology and Operative Intensive Care Medicine, University Hospital Augsburg, Stenglinstrasse 2, 86156 Augsburg, Germany
[10]Department of Technology, Management, and Economics, Technical University of Denmark
[11]Data and Development Support, Region Zealand, Denmark
[12]Professor of Applied Data Science in Health Care, Nürnberg School of Health, Ohm University of Applied Sciences Nuremberg

Correspondence:
milena.grieger@uni-a.de

Submission: November 2023

Original Research Paper:

***DENLU* and *leaky stanh*: customized activation functions targeting enhanced sensitivity with healthcare applications in binary classification**

Abstract. In healthcare settings, binary classification of patients is frequently applied. For example, hospital decision-makers seek to classify patients regarding a positive or negative diagnosis and intensive care unit (ICU) treatment or non-ICU, respectively. Lately, the importance of machine learning algorithms as decision support tools for the classification of patients has significantly risen in literature. While most machine learning applications consider both classes equally important, the focus in healthcare is often on one particular class, e.g., ICU treatment and enhanced sensitivity for this class. New customized activation functions (AFs), namely *Double Exponential Non-Linear Unit (DENLU)* and *Leaky Scaled Hyperbolic Tangent (leaky stanh)*, with applications in binary classification of healthcare data, are studied. The performance of the functions is tested and compared with existing ones for simulated and real-world healthcare data. The results show that the machine learning model's performance based on the customized AFs is superior to existing ones in terms of sensitivity (up to +17.7 percentage points) and AUC (up to +7.6 percentage points) but depends strongly on the heterogeneity of the data. From a theoretical perspective, *DENLU* and *leaky stanh* show interesting properties concerning the S-shape flexibility and the avoidance of the problem of vanishing gradients. Thus, the customized AFs may be used for binary classification of heterogeneous healthcare data if a focus on sensitivity is required. Accurate predictions for a specific class enable a decision maker to formulate more exact plans, e.g., the necessity of ICU treatment.

Keywords: medical decision making; machine learning; binary classification; activation function; sensitivity

1 Introduction

In healthcare settings, binary classification of patients is frequently applied. For example, healthcare decision-makers seek to classify patients regarding a positive or negative diagnosis, in- or outpatients, and intensive care unit (ICU) treatment or non-ICU treatment, respectively. The classifications are essential for effective and efficient hospital resource planning and contribute to optimal medical care. If, for example, a patient is mistakenly transferred from the emergency department to the ward when he or she should be receiving ICU treatment, this leads to a strain on resources and, at the same time, endangers the patient's health. Lately and probably fueled by the Covid-19 pandemic, the importance of machine learning algorithms as decision support tools has significantly increased in literature (for example: Alballa and Al-Turaiki (2021); Pfeuffer et al. (2023); Weber et al. (2022); Wynants et al. (2020)). While most machine learning applications consider both classes equally important, the focus in healthcare is often on one particular class, e.g., ICU treatment.

Furthermore, there exist solely a very limited number of activation functions (AFs) that can be integrated into the algorithms. Commonly used are the *sigmoid, Exponential Linear Unit (ELU), Rectified Linear Unit (ReLU),* and *Hyperbolic Tangent (tanh)* functions (Ohn and Kim 2019). *Sigmoid* and *tanh* have formed the basis for a large part of the AFs used in the field of healthcare and medicine for quite a long time (El-Baz and Suri 2021). While both meet the criteria of the Universal Approximation Theorem (UAT) for AFs and show, other than *ELU* and *ReLU,* a high training stability due to their S-shape, *sigmoid* is not flexible in scaling the S-curve. A significant limitation of all AFs mentioned is generally known as the vanishing gradient. Vanishing gradients describe gradients becoming continuously smaller and converging to 0 with increasing absolute values of the independent variable (Li et al. 2014). As *ELU* and *ReLU* are non-S-shaped curves, they are subject to 'one-sided' vanishing gradients problem only.

In this work, we propose customized AFs in machine learning, namely *Double Exponential Non-Linear Unit (DENLU)* and *Leaky Scaled Hyperbolic Tangent (leaky stanh)*, with applications in binary classification of healthcare and medicine data. Both are extended versions of well-known AFs. The former is a flexible S-shaped alternative to *sigmoid* based on *(R)ELU*. The latter is an extension of *(s)tanh* solving the problem of vanishing gradients. Both functions are based on the criteria defined by the UAT for AFs. We want to gain a better understanding of how AFs based on UAT can impact the outcomes of a machine learning model. Their performance is tested for simulated and real-world healthcare data. The data includes three different data sets on Covid-19 diagnosis based on laboratory parameters, classification of Covid-19 patients in the emergency department regarding ICU treatment, and classification of elective patients regarding ward or ICU treatment after surgery. Medical decision-makers need to focus on highly sensitive classifications in all these real-world scenarios. For example, individuals who are misclassified as Covid-19 negative are likely to become additional spreaders of the disease if not appropriately isolated. Moreover, ICU capacity belongs to one of the scarcest resources in hospitals. Thus, if patients suffering from Covid-19 or undergoing surgery are wrongly predicted not to need ICU treatment, capacity constraints could cause severe health issues for those individuals. Therefore, all use cases mentioned above benefit from highly sensitive results, so the sensitivity of the AFs is considered explicitly in the remainder of this paper. Moreover, we examine how homogeneous and heterogeneous data can influence the outcomes of machine learning models, aiming to glean new insights within the context of data structure.

Our work is structured as follows. In section 2, we provide a review of existing AFs. The review is the basis for introducing our customized AFs, *DENLU* and *leaky stanh*, in section 3. The AFs are applied to simulated and real-world healthcare data, and the results are discussed in detail in section 4. Section 5 concludes and provides an outlook for future research.

2 A review of activation functions

Despite the considerable significance of selecting an appropriate AF, there is no standardized protocol within the research for guiding the AF selection or development process. This lack of consensus can be attributed, in part, to the diverse range of prediction tasks to which AFs can be applied to varying degrees (Liew et al., 2016). Nevertheless, one potential solution to this issue is to use the UAT as a guideline when creating an AF. As a result, AFs have the desired property of approximating an arbitrary, nonlinear (Hornik 1991), continuous function with arbitrary accuracy (Hornik et al. 1989). For this purpose, AFs must meet specific criteria, specifically the Universal Approximation Properties (UAP) underlying the UAT, listed in Table 1 (Sodhi and Chandra 2014). Given the UAT, various methods can be employed to categorize AFs.

UAP.1	AF $\sigma(x)$ creates output that is non-constant for the entire input range.
UAP.2	AF is bounded within a value domain with P as an upper bound and $\lvert\sigma(x) \leq P\rvert$ holds.
UAP.3	AF is continuous over all input values c, so that $\lim_{x \to c} \sigma(x) = \sigma(c)$ holds.
UAP.4	AF is a monotonically increasing function, so that $\sigma(x) \leq \sigma(y)$ for $x \leq y$ is valid.
UAP.5	AF is differentiable everywhere (i. e., identical slope of the AF for each input value for both a left-sided and a right-sided convergence to this value).

Table 1: Universal approximation properties

Sigmoidal vs. non-sigmoidal functions. A frequently chosen variant is to categorize AFs based on their curve shapes. Within this framework, the functions are divided into a group of so-called sigmoidal or non-sigmoidal functions (Chandra et al. 2015). The former group includes functions that create outputs shaped as S-curves, whereas the latter group does not exhibit this characteristic. The *logistic (sigmoid) function* $\sigma(x) = \frac{1}{1+e^{-x}}$ is a representative example of the former group, and the *ReLU* $\sigma(x) = \max(x, 0)$ is the most prevalent AF in the latter group (Liew et al. 2016).

Piecewise linear vs. locally quadratic functions. Another option for categorizing AFs is classifying them into piecewise linear and locally quadratic functions. Piecewise linear functions have no curvature for certain intervals which are separated at breakpoints k and l, i.e., the corresponding function

section is linear. Here, k and l are used as two exemplary breakpoints. Consequently, the first-order derivative is constant in such an interval $[k, l]$, i.e., the second-order derivative is zero. A well-known representative of the piecewise linear AFs is ReLU and its parameterized modification leaky ReLU $\sigma(x) = \max(x, \alpha x)$. Locally quadratic functions include functions with at least one open interval (a, b) with a non-zero second-order derivate. A further distinction within locally quadratic functions involves smooth functions with non-zero second-order derivatives throughout their domain, such as the sigmoid function. The remaining functions are called piecewise smooth. ELU as a piecewise smooth function, and the tanh as a smooth function can be mentioned as exemplary representatives (Ohn and Kim 2019). In analogy to ReLU, ELU can be expressed in a parameterized form:

$$\sigma(x) = \begin{cases} \alpha(e^x - 1) & \text{if } x \leq 0 \\ x & \text{else} \end{cases} \tag{1}$$

Moreover, a parameterized representation is also possible for *tanh* by introducing the amplitude a as well as the slope b, resulting in the *Scaled Hyperbolic Tangent (stanh) (Liew et al. 2016)*:

$$\sigma(x) = a \cdot \tanh(bx) = a \cdot \frac{e^{bx} - e^{-bx}}{e^{bx} + e^{-bx}} \tag{2}$$

Other approaches. Lastly, we provide a concise overview of another category of AFs that differs fundamentally from the previously mentioned division into distinct groups. Until now, the focus has primarily been on empirically fixed AFs, with only the model parameters being updated through the backpropagation algorithm. Notably, the AF itself remains unaltered during the training process and is, as previously mentioned, empirically determined. In contrast to this prevailing approach, an alternative method involves approximating a conventional AF $\sigma(x)$ by employing a Taylor series expansion. For a more detailed exploration of this unconventional approach, we refer to the to the expositions of Chung et al. (2016).

3 An introduction to *DENLU* and *leaky stanh*

In the following, the customized AFs *DENLU* and *leaky stanh* are presented based on the properties of several functions listed before. For this purpose, we modify some existing functions, namely *ELU* and *stanh*, so that the requirements according to the UAT are fulfilled as extensively as possible. These modifications aim to reduce the shortcomings of already known functions, such as the vanishing gradient, shape inflexibility, and training instability (see below). Additionally, we seek to improve the sensitivity of the AFs in comparison with their original basis and the respective use cases.

3.1 *DENLU*

The first AF we propose is *DENLU*. According to its name, the *DENLU* function is derived from the *ELU* function introduced before. In contrast to *ELU*, *DENLU* includes both a left-sided and a right-sided saturation behavior. In this context, it is worth mentioning that the upper and lower bounds of the value domain and, hence the saturation behavior of an AF significantly impact the training stability (Puheim et al. 2014). Further limiting the value domain as represented by the two-sided saturation is to reduce the training instability to which the modifications of the *ReLU* function, including *ELU*, are sensitive (Liew et al. 2016). Analogous to *ELU*, there is a left-sided saturation for $x < 0$. In addition, for $x > 0$, a right-sided saturation is introduced with a reversed curvature behavior compared to $x < 0$, with the origin representing an inflection point. Due to the two-sided saturation, the original linear function section of the *ELU* function is omitted for *DENLU*. Therefore, the definition of S-shaped *DENLU* is as follows:

$$\sigma_1(x) = \begin{cases} e^x - 1 & \text{if } x \leq 0 \\ 1 - e^{-x} & \text{else} \end{cases} \tag{3}$$

With the UAT in mind, it is worth mentioning that the *DENLU* function is derivable for all values. While the *ELU* function has the value domain $[-1, \infty]$, *DENLU* ranges outputs between -1 and 1, as seen in Figure 1.

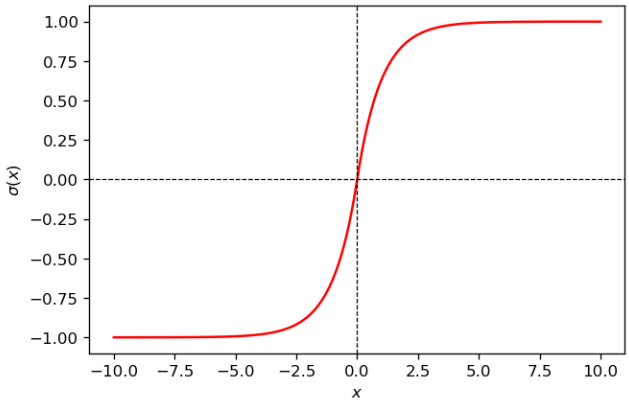

Figure 1: Double Exponential Non-Linear Unit (DENLU)

As is the case for the regular *ELU* function, *DENLU* can be extended as a parameterized function using the parameter γ resulting in an adjusted value domain $[-\gamma, +\gamma]$. Thus, the parameterized *DENLU* function can be formulated in the following way:

$$\sigma_1(x) = \begin{cases} \gamma \cdot (e^x - 1) & \text{if } x \leq 0 \\ \gamma \cdot (1 - e^{-x}) & \text{else} \end{cases} \tag{4}$$

How *DENLU* changes for different values of γ can be seen in Figure 8 in the Appendix. Given that *DENLU* is derived based on *ELU*, the difference between these AFs is worth highlighting. *DENLU* differs from *ELU* because for $x > 0$, *DENLU* is $1 - e^{-x}$ and *ELU* is x. As indicated at the beginning of this section, this bounded range of values of *DENLU* for both $x \leq 0$ and $x > 0$ and thus the resulting two-sided saturation aims to improve the training stability. While looking at *ELU* the difference to *DENLU* is readily apparent, the difference from *DENLU* to *stanh* is not immediately obvious since both *DENLU* and *stanh* fulfill the same UAPs. To make this clearer, a closer look at the value ranges of the two functions is necessary. For illustrative purposes, the hyperparameters α (*DENLU*), a and b (*stanh*) are assumed to be 1 each. On the one hand, for $x > 0$, *DENLU* differs from *stanh* by $-\frac{(1-e^{-x})^2}{e^x+e^{-x}}$, i.e., *DENLU* is smaller than *stanh* by $\frac{(1-e^{-x})^2}{e^x+e^{-x}}$ in this interval. On the other hand, *DENLU* is larger than *stanh* by $\frac{(1-e^{x})^2}{e^x+e^{-x}}$ for $x \leq 0$. This means that the absolute value of

DENLU is smaller by $\frac{(1-e^x)^2}{e^x+e^{-x}}$ and thus less negative in the negative domain of the function. It follows that *DENLU* is smaller in absolute value than stanh in every subdomain of the function except for $x = 0$, which again indicates an increased training stability. Given that α and a represent equivalent hyperparameters for *DENLU* and *stanh*, respectively, the value domain of *stanh* can be further modified via b, however, this requires an additional parameter to be determined. For a detailed derivation of the differences between *DENLU* and *stanh*, we refer the reader to the Appendix.

3.2 *Leaky stanh*

The second AF we propose is *leaky stanh*. This AF represents a modification of the function *stanh,* which intends to prevent the problem of the so-called vanishing gradient to which sigmoidal functions are subject. As it is already well-known, the parameters of a neural network are updated by the backpropagation of error gradients using the chain rule during the training process. A problem that arises with sigmoidal functions is that due to the increasing saturation of these functions for $x \longrightarrow \infty$ and $x \longrightarrow -\infty$ those gradients become continuously smaller and converge to 0 as network parameters are updated over the network layers (Li et al. 2014). This problem generally known as vanishing gradient is also referred to as gradient diffusion and might negatively affect training performance since common learning algorithms rely on these gradients when optimizing network parameters (Liew et al. 2016).

Following the theoretical considerations mentioned above, a modification of the *stanh* function is introduced in such a way that its margins are replaced each by a linear function section. The slope of these function sections represents an additional parameter of the customized AF. Accordingly, the customized AF is composed of three sections, consisting of the original *stanh* for $c \leq x \leq d$ and the two linear function sections at the margins $x < c$ and $x > d$. Consequently, the three sections are combined at positions c and d to form the *leaky stanh* function. In accordance with UAP.5, joining these sections

must be executed in such a way that the resulting function is differentiable everywhere. Therefore, the sections need to be combined at the specific positions c and d, respectively, where the respective function sections under consideration show the same slope. To ensure this, the first order derivative (FOD) of the Scaled Hyperbolic Tangent is considered as shown in Figure 2.

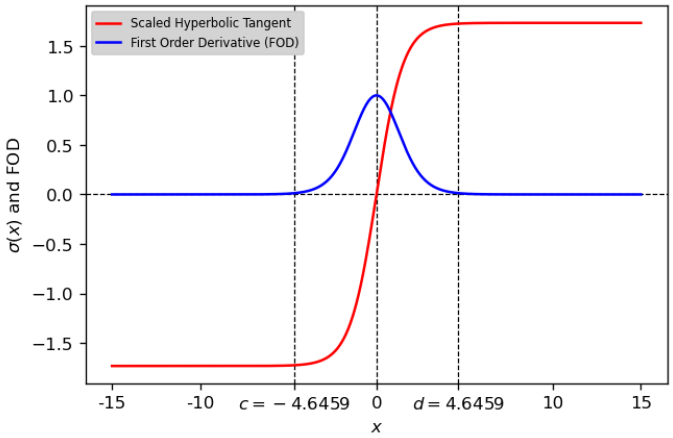

Figure 2: Scaled Hyperbolic Tangent and FOD

Here, c and d are located at the positions at which the FOD, i.e., the gradient, of *stanh* corresponds to the slope of the linear function sections. This slope hereafter referred to as γ, is a parameter and can be specified in advance. In the following, γ is exemplarily set to 0.01 for illustrative purposes. In addition, according to recommendations in literature, the parameters $a = 1.7321$ and $b = 0.6585$ are set (Liew et al. 2016). However, both a and b, as well as γ, can be varied. The determination of c and d, given in Figure 2, can be shown as follows.

The first step is to determine the FOD of the *stanh* that is required to correspond to the slope of the linear function:

$$\frac{\partial}{\partial x} a \cdot \tanh(bx) = \frac{\partial}{\partial x} a \cdot \frac{e^{bx} - e^{-bx}}{e^{bx} + e^{-bx}} = \frac{4abe^{2bx}}{(e^{2bx} + 1)^2} \stackrel{!}{=} \gamma \tag{5}$$

The next step is to solve equation (5) for x such that c and d are obtained as solutions x_1 and x_2:

$$\frac{4ab}{\gamma} - 2 = e^{2bx} + e^{-2bx} \tag{6}$$

$$x = \frac{\ln\left\{\frac{1}{2}\left(\left(\frac{4ab}{\gamma} - 2\right) \pm \sqrt{\left(\frac{4ab}{\gamma} - 2\right)^2 - 4}\right)\right\}}{2b} \tag{7}$$

Thus, according to (7), c and d are given by:

$$c = x_1 = \frac{\ln\left\{\frac{1}{2}\left(\left(\frac{4ab}{\gamma} - 2\right) + \sqrt{\left(\frac{4ab}{\gamma} - 2\right)^2 - 4}\right)\right\}}{2b} \tag{8}$$

$$d = x_2 = \frac{\ln\left\{\frac{1}{2}\left(\left(\frac{4ab}{\gamma} - 2\right) - \sqrt{\left(\frac{4ab}{\gamma} - 2\right)^2 - 4}\right)\right\}}{2b} \tag{9}$$

As shown in Figure 2, using $a = 1.7321$, $b = 0.6585$ and $\gamma = 0.01$ given as examples above, $c \approx -4.6459$ and $d \approx 4.6459$ are obtained. Subsequently, based on amplitude a, slope b, function section slope γ as well as differentiation parameters c and d, the proposed AF *leaky stanh* can be formulated as follows:

$$\sigma_2(x) = \begin{cases} a \cdot \tanh(bc) + \gamma \cdot (x - c) & \text{if } x < c \\ a \cdot \tanh(bx) & \text{if } c \leq x \leq d \\ a \cdot \tanh(bd) + \gamma \cdot (x - d) & \text{if } x > d \end{cases} \tag{10}$$

Equation (10) shows that $\sigma_2(x)$ is a linear function with slope γ for $x < c$ and $x > d$. In this context, the constants $a \cdot \tanh(bc)$ and $a \cdot \tanh(bd)$ denote the intercept of the respective linear section. As already mentioned at the beginning of this section, σ_2 is equivalent to the *stanh* for $c \leq x \leq d$. Again, considering the exemplary values for $a = 1.7321$, $b = 0.6585$ as well as $c \approx -4.6459$ and $d \approx 4.6459$ obtained from (8) and (9), (10) can be reformulated:

$$\sigma_2(x) = \begin{cases} -1.7245 + \gamma \cdot (x + 4.6459) & \text{if } x < -4.6459 \\ 1.7321 \cdot \tanh(0.6585x) & \text{if } -4.6459 \leq x \leq 4.6459 \\ 1.7245 + \gamma \cdot (x - 4.6459) & \text{if } x > 4.6459 \end{cases} \quad (11)$$

Equation (11) is graphically illustrated in Figure 3.

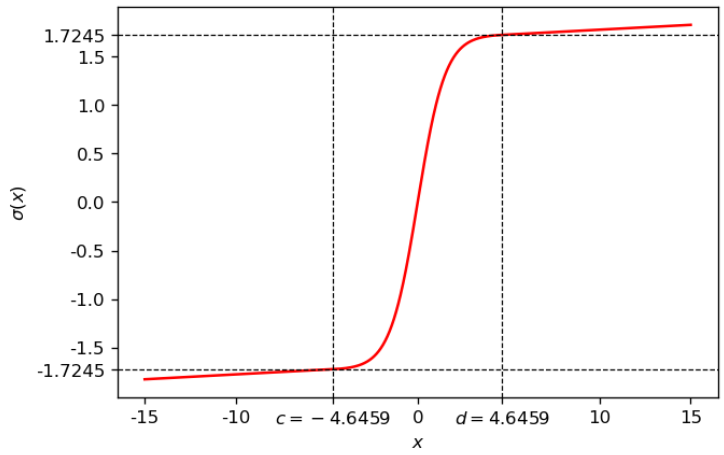

Figure 3: Leaky scaled hyperbolic tangent (leaky stanh)

How *leaky stanh* changes for different values of a and b can be seen in Figure 9 in the Appendix.

Summarizing, *leaky stanh* belongs to the category of sigmoidal and piecewise linear functions, while *DENLU* belongs to the category of sigmoidal and locally quadratic functions. *DENLU* fulfills all properties of the UAT, while *leaky stanh* does not satisfy the second property of the UAT in certain cases. *DENLU* and *leaky stanh* have flexibility in shape scaling and ensure training stability, unlike *ELU* and *ReLU*, which suffer from the 'one-sided' vanishing gradient problem. Table 2 provides an overview of the categorization and properties of *DENLU* and *leaky stanh* compared to *ELU*, *tanh*, *ReLU*, *sigmoid,* and *stanh* functions.

Category/Property	***DENLU***	*ELU*	*tanh*	***leaky stanh***	*ReLU*	*sigmoid*	*stanh*
Sigmoidal	✓		✓	✓		✓	✓
Non-sigmoidal		✓			✓		
Piecewise linear				✓	✓		
Locally quadratic	✓	✓	✓			✓	✓
UAP.1	✓	✓	✓	✓		✓	✓
UAP.2	✓		✓			✓	✓
UAP.3	✓	✓	✓	✓	✓	✓	✓
UAP.4	✓	✓	✓	✓	✓	✓	✓
UAP.5	✓	(✓)	✓	✓		✓	✓
Shape flexibility	✓	✓		✓	✓		✓
Training stability	✓		✓	✓		✓	✓
No vanishing gradient		(✓)		✓	(✓)		

Table 2: Comparison of the categorization and properties of existing AFs, DENLU and leaky stanh

4 Application to simulated and real-world healthcare data

In applying the two AFs *DENLU* and *leaky stanh*, we focus on binary classification of patients in a healthcare context. For this purpose, we apply a machine learning algorithm to both simulated and real-world healthcare data using the TensorFlow library with the Python programming language. In the following, we define our machine learning algorithm and the performance measures, introduce the data sets and data preparation, and present and discuss the results.

4.1 The machine learning algorithm and performance measurement

For better comparability of the AFs, the same deep neural network is used for all data sets. Since the problem of the vanishing gradient occurs mainly with deep neural networks, a neural network with more than one hidden layer is considered in this application. In the hidden layers, we use the AF under consideration (e.g., *DENLU*) while we use *sigmoid* in the output layer. The latter represents one of the two most applied AFs in healthcare settings. By following this procedure, we ensure comparability between our obtained results.

A binary cross-entropy is applied as a loss function, and gradient descent with a learning rate of 0.01 defines the optimizer. The settings are derived from extensive pre-testing. To measure the performance, we consider accuracy, sensitivity, specificity, F1-Score, area under the curve (AUC), precision (positive predicted value, PPV), and negative predicted value (NPV). We compare *DENLU* and *leaky stanh* with *ELU*, *ReLU*, *tanh,* and *sigmoid*. Ten-fold cross-validation is applied to avoid overfitting. The algorithm's training includes 85 epochs with a batch size of 10. Hyperparameter tuning was deferred to a second step specifically because applying it beforehand would have rendered the results incomparable.

4.2 Data sets and data preparation

The machine learning algorithm is applied to simulated data, Covid-19 triage data, Covid-19 diagnosis data and elective surgery and intensive care (ESIC) data. Since healthcare data often deals with common problems like missing values or unbalanced classes (Das et al. 2019), extensive data preparation including feature scaling was necessary. Table 3 provides a detailed overview of the data sets and data preparation applied to the data sets. The training-test-split for all data sets is 90 % to 10 %.

Data set	**Simulated**	**Covid-19 triage**	**Covid-19 diagnosis**	**ESIC**
No. of patients	10,000	3,543	3,670	26,600
No. of features	20	59	14	12
No. of classes	2	2	2	2
Positive class	Class 1	ICU treatment	SARS-CoV-2 pos.	ICU treatment
Negative class	Class 0	Non-ICU treatment	SARS CoV-2 neg.	Non-ICU treatment
Feature scaling	✓	✓	✓	✓
Feature heterogeneity	✓	✓		✓
Class balance by	Data generation	Oversampling	Oversampling	Oversampling
Missing values filled by	-	Random Forest	Random Forest	Multi-Layer Perceptron

Table 3: Overview of the data sets and data preparation applied to the data sets

Simulated data. Our simulated data set consists of a $10{,}000 \times 20$ multivariate normal $N(\mu; \sigma)$ feature matrix with expectation vector μ and covariance matrix σ as well as a $10{,}000 \times 1$ binary label vector. Regarding the distribution of labels, the data set is balanced, i.e., $5{,}000 \times 20$ random numbers belong to the first class ($cl = 0$) and $5{,}000 \times 20$ random numbers belong to the second class ($cl = 1$). The positive definite covariance matrix σ is of 20×20 type with principal diagonal elements $\sigma_{ii} = 1$ for all $i = 1, \dots, 20$ and a randomly generated dependency structure σ_{ij} for all $i, j = 1, \dots, 20$ with $i \neq j$. While the covariance structure is the same for all $10{,}000 \times 20$ random numbers in the data set, the components of the expectation vector vary depending on the class and are defined as follows:

$$\mu_{i_{cl}} = \begin{cases} UNIF_0[0,1] & \text{for all } i = 1, \dots, 20 \text{ and } cl = 0 \\ UNIF_1[0,1] & \text{for all } i = 1, \dots, 20 \text{ and } cl = 1 \end{cases} \tag{12}$$

Covid-19 triage data. The data set focuses patients tested positive for SARS-CoV-2 from March 18, 2020, to January 7, 2021, and is based on a Lean European Open Survey on SARS-COV-2 Infected Patients (LEOSS) export. LEOSS is a European multicenter cohort study enabling extensive retrospective data analyses. Based on 3,543 cases and 59 documented features (see Table 4), which include, e.g., vital signs or laboratory values, the machine learning algorithm is used to predict whether a patient arriving in the emergency department will require an ICU bed. This AI-based classification of incoming patients might support physicians in their decision making. Features with underfilled columns in the data set are removed and comorbidities are combined using the Charlson Comorbidity Index (CCI, (Charlson et al. 1987)). In general, features in the data set are rather heterogeneous, because different categories such as vital signs, lab values, imaging outcomes, and symptoms are included.

General information	Symptoms	Vital signs	Lab values	Imaging
• At least one neuronal disease	• Non-specific Covid-19 symptoms	• Systolic blood pressure	• Aspartate transaminase	• CT: Air trapping
• At least one cardiovascular disease	• Other neurological findings	• Diastolic blood pressure	• Alanine transaminase	• CT: Areas of consolidation
• Gender	• Sore throat	• Pulse	• Creatinine	• CT: Bronchiolitis
• Age	• Dry cough	• Temperature	• Bilirubin	• CT: Crazy paving pattern
• Prior heart failure			• Gamma-glutamyl transferase	• CT: Ground glass opacities
• Stage of heart failure	• Productive cough	• Respiratory Rate	• Urea	• CT: Interlobular septal thickening
• BMI	• Wheezing	• Oxygen saturation	• Lactate dehydrogenase	• CT: Nodular lesions
• CCI / Sum of comorbidities	• Dyspnoea		• D-dimer	• CT: Pleural effusion
	• Palpitations		• Leukocytes	• Other relevant CT results
	• Diarrhea		• Lymphocytes	
	• Muscle aches		• Neutrophils	
	• Muscle weakness		• Platelets	
	• Fever		• Hemoglobin	
	• Delirium			
	• Excessive tiredness			
	• Headache			
	• Meningism			
	• Smell disorder			
	• Taste disorder			
	• Runny nose			
	• Red eye			

Table 4: Features of the Covid-19 triage data set

Covid-19 diagnosis data. In addition to Covid-19 triage data, we use Covid-19 diagnosis data. Our data origins from three different sources (University Hospital of Augsburg, Germany, Alb-Fils Kliniken Göppingen, Germany, and LEOSS registry). All data has been collected during first pandemic wave in Germany in 2020. The outcome of the machine learning algorithm is to classify whether a symptomatic patient is infected with SARS-CoV-2. From an organizational perspective, a digital Covid-19 diagnosis might accelerate the processes in the emergency department, for example, by providing a fast alternative to Nucleid Acid Amplification based tests. The full data set consists of 3,670 patients. After preprocessing, the data set provides 14 features per

patient, while the features are rather homogeneous including information on lab values, age, and gender (see Table 5).

General information	Lab values
• Age • Sex	• C-reactive protein • Direct bilirubin • Erythroblasts • Gamma-glutamyl transferase • Hemoglobin • Leukocytes • Partial thromboplastin time • Platelets • Serum alanine transaminase • Serum creatinin • Serum lactate dehydrogenase • Serum urea

Table 5: Features of the Covid-19 diagnosis data set

ESIC data. Besides predictions on the Covid-19 pandemic, our extended AFs are tested for another healthcare scenario. Since ICU beds are scarce resources in hospitals, the decision on a patient's elective surgery includes the question of whether this patient needs ICU treatment after surgery. To digitally support the decision process of physicians, we use a data set from the University Hospital of Augsburg, Germany. The data set consists of a selection of 26,600 surgeries that were conducted between 2017 and 2021. The 12 heterogeneous features in the data set belong to three categories: characteristics of the respective surgery, metrics about the patient's physical appearance, and indicators about the patient's general health circumstances, which are listed in detail in Table 6. Additionally, every surgery has a binary label on whether the patient went to the ICU after elective surgery.

Surgical characteristics	Patient's physical appearance	Patient's health circumstances
• Medical specialty • Estimated surgery duration • Estimated anesthesia duration • Type of anesthesia • Main diagnosis for a hospital stay	• Age • Sex • Weight • Height	• Number of comorbidity diagnoses • CCI • ASA-score

Table 6: Features of ESIC data set

4.3 Results

The results section starts by examining the individual data sets. Subsequently, a sensitivity analysis and a comparison of the results are discussed. This part focuses on the performance measures accuracy, AUC, and sensitivity because of their relevance to the data sets and application area. Accuracy is the standard measure for machine learning models, AUC is a combination of other measures (sensitivity and false positive rate), and sensitivity, as motivated above, is particularly important in our application domain. The ranks per data set and AF for the performance measures accuracy, AUC, and sensitivity are shown in Table *7*. Detailed results, including all performance measures, can be found in Table 8 in the Appendix. In the following, absolute values are expressed as % and differences in percentage points (PP).

Covid-19 triage data. For the Covid-19 triage data set, our proposed AFs show the following performance: For *DENLU* (*leaky stanh*), the average performance across all folds for the test data set is 72.65 % (72.40 %) for accuracy, 69.94 % (69.51 %) for sensitivity, and 79.49 % (79.05 %) for AUC. For the benchmark AFs, *sigmoid* achieves an average accuracy of 73.24 %, *ELU* of 71.94 %, *tanh* of 71.15 %, and *ReLU* of 71.72 %. Thus, compared to *DENLU (leaky stanh*), only *sigmoid* achieves better accuracy of +0.9 (+0.6) PP. These findings can also be obtained with the performance measure AUC. *DENLU (leaky stanh)* achieves almost the same score of −0.04 (−0.05) PP compared to *sigmoid*, but a better score of up to +7.6 (+7.2) PP compared to *tanh*, *ReLU,* and *ELU*. For *DENLU (leaky stanh)*, a deviation from +14.3 (+13.9) PP to +17.7 (+17.3) PP is possible for sensitivity compared to the other AFs *tanh*, *ReLU,* and *ELU* in the test data set. Only *sigmoid* achieves slightly better sensitivity than *DENLU* (*leaky stanh*) (+1.2 (+1.7) PP). Thus, *DENLU*, *leaky stanh,* and *sigmoid* show superior performance for enhanced sensitivity in this data set. The resulting performance of the different AFs can be seen in Figure *4*.

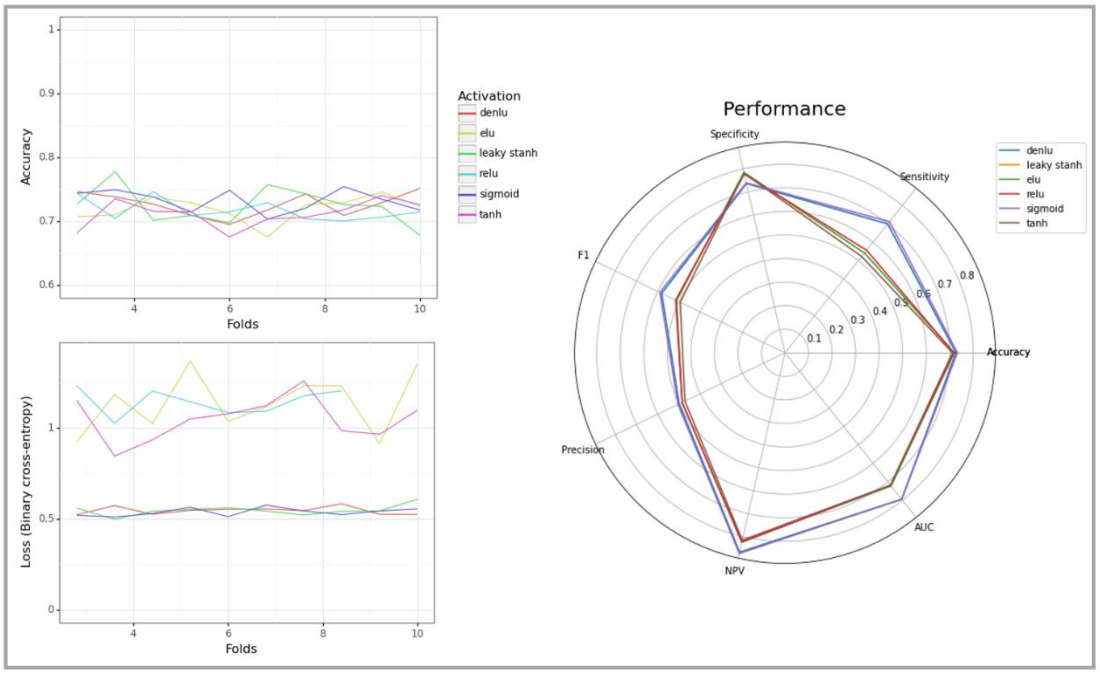

Figure 4: Performance of AFs (Covid-19 triage test data)

As *DENLU* and *leaky stanh* achieve similar results on average over the folds, they cannot be clearly distinguished in the radar chart. Evaluation of the ranks highlights the effects that *sigmoid*, *DENLU*, and *leaky stanh* perform best for the Covid-19 triage data set (see Table 7).

Covid-19 diagnosis data. Concerning the Covid-19 diagnosis data set, the two AFs presented perform as follows. The average accuracy for the test data set is 82.26 % (82.34 %) for *DENLU* (*leaky stanh*), sensitivity is 81.74 % (81.80 %), and AUC is 89.82 % (89.97 %). When considering the AFs to be compared, for example, sensitivity is higher for *sigmoid* at 83.14 %, and considerably higher for *ELU* and *ReLU* and *tanh* at approximately 91 %, 92 %, respectively 92 %. For accuracy, the difference between the existing AFs and *DENLU* (*leaky stanh*) is up to −8.9 (−8.8) PP. The results regarding accuracy, loss, and the overall comparison of the performance of the AFs are shown in Figure 5.

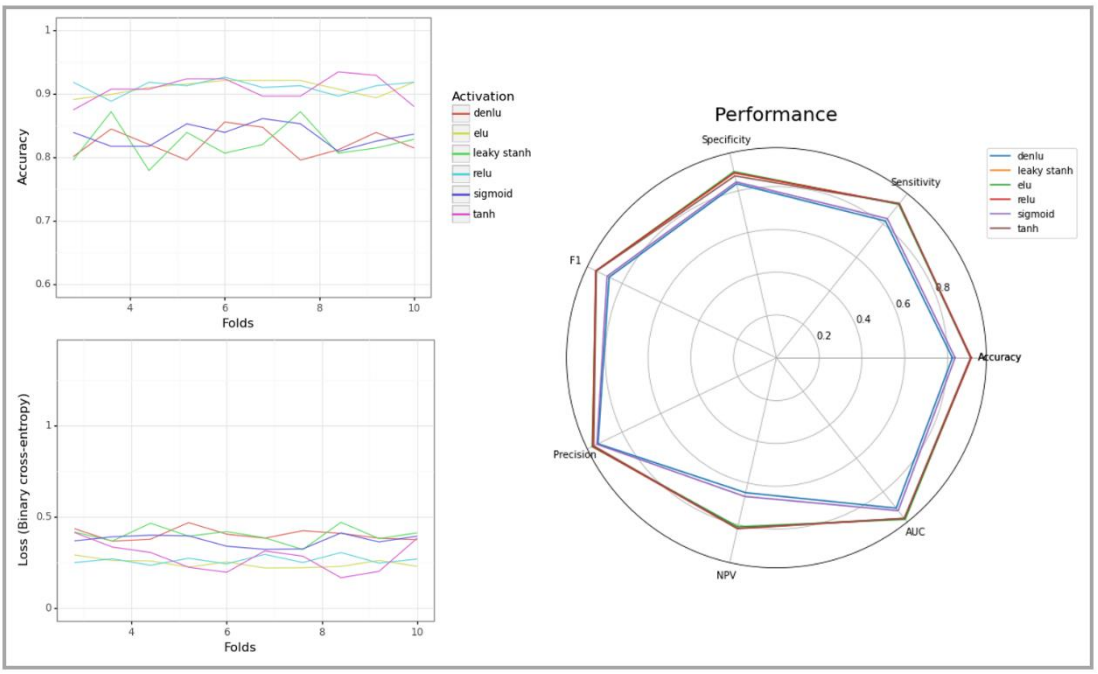

Figure 5: Performance of AFs (Covid-19 diagnosis test data)

ESIC data. For the ESIC test data, the performance measures accuracy, sensitivity, and AUC are 86.70 % (86.97 %), 86.43 % (86.33 %), and 92.94 % (92.87%) for *DENLU* (*leaky stanh*). Compared to existing AFs, the accuracy is lower (up to −2.5 PP), but the AUC is higher (up to +2.9 PP). The difference in *DENLU* (*leaky stanh*) is up to +9.6 (+9.5) PP for sensitivity compared to *ELU, ReLU, sigmoid,* and *tanh*. As before, in this data set, the lower accuracy of *DENLU* and *leaky stanh* can be compensated with better performance in terms of AUC and significantly better performance in terms of sensitivity. The results for the ESIC data set are summarized in Figure *6*. The evaluation across the ranks in this data set also shows a leady role for *DENLU* and *leaky stanh* (see Table 7).

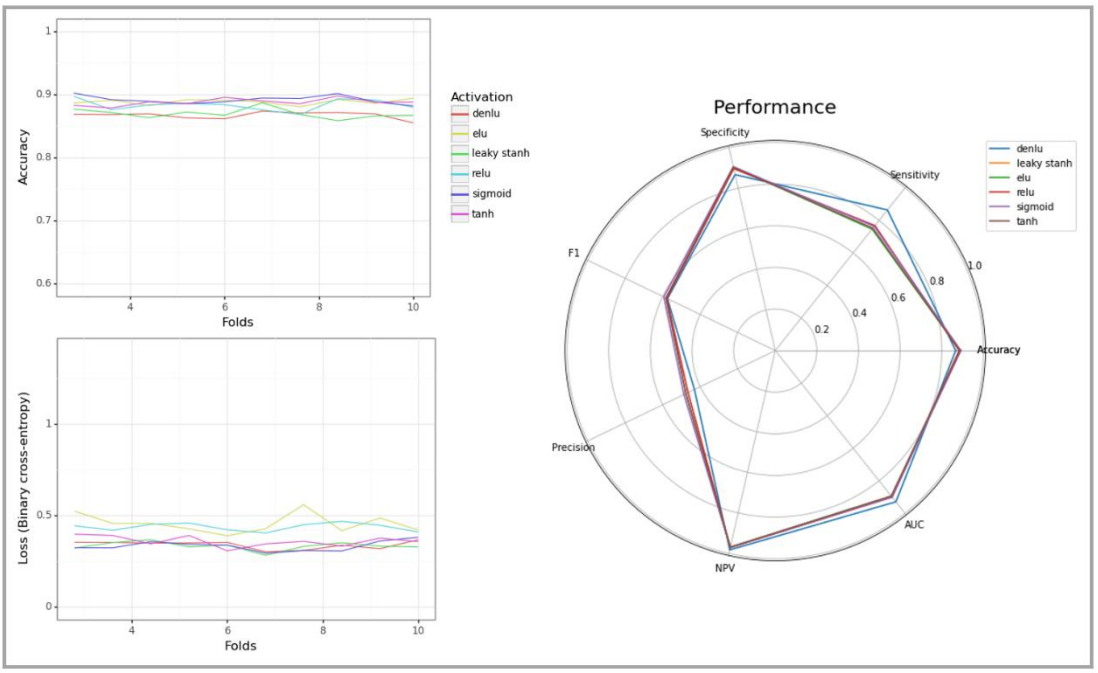

Figure 6: Performance of AFs (ESIC test data)

4.3.1 Hyperparameter tuning and sensitivity analysis

To find a suitable neural network for the respective data set, hyperparameter tuning was performed for the two customized AFs. The considered hyperparameters include the number of hidden layers (1 to 5), the number of neurons within the hidden layer (10 to 500 in steps of 10), the AF in the output layer (*sigmoid* and *softmax*), and the learning rate (0.0001 to 0.01). The individual neural networks are presented in Table 9 in the Appendix. Regarding the simulated test data set, the two AFs *DENLU* and *leaky stanh* perform relatively similarly with an accuracy up to 94.45 %. The performance regarding sensitivity drops down by −0.9 PP (from 95.60 %). The improved neural network for Covid-19 triage data with one hidden layer for *DENLU* and three hidden layers for *leaky stanh* achieves much better performance in terms of accuracy. The accuracy increases in the test data set for *DENLU* by +7.6 PP and for *leaky stanh* by +5.7 PP, but sensitivity decreases to a minimum of 55.56 %. Thanks to hyperparameter tuning, the performance of *DENLU* and *leaky stanh* can be improved by up to +3.1 PP in terms of accuracy and up to +5.4 PP in terms of sensitivity for the Covid-19 diagnosis data. For the ESIC

data set, hyperparameter tuning also improves the performance of the accuracy of *DENLU* and *leaky stanh* by up to +2.4 PP but worsens the sensitivity by −5.5 PP. In addition, single-criteria sensitivity analyses were conducted to examine the behavior of *DENLU* and *leaky stanh* by altering the respective parameters. For *DENLU*, varying the parameter γ from 0 to 1 in steps of 0.2 showed almost no change in performance measures across all data sets. For instance, the sensitivity (accuracy) of the simulated data changes by a maximum of +0.3 (+0.3) PP and Covid-19 triage data by a maximum of +1.8 (+0.5) PP. Similarly, for *leaky stanh*, altering parameters γ (from 0.1 to 0.4), a (from 1.6 to 1.8), and for b (from 0.5 to 0.8) in steps of 0.1 each. Again, we have found no significant changes in the average values of the performance measures (e.g., sensitivity (accuracy) changes by a maximum of +1.5 (+0.7) PP). However, these findings are specific to the tested data sets and should be verified for other data sets individually.

Data set	Measure	***DENLU***	***leaky stanh***	*ELU*	*ReLU*	*sigmoid*	*tanh*
Simulated	Accuracy	*1*	*2*	*4*	*5*	*3*	*6*
	AUC	*1*	*2*	*4*	*5*	*3*	*6*
	Sensitivity	*2*	*1*	*6*	*3*	*5*	*4*
	Rank	***1 (1.33)***	***2 (1.67)***	***5 (4.67)***	***4 (4.33)***	***3 (3.67)***	***6 (5.33)***
Covid-19 triage	Accuracy	*2*	*3*	*4*	*5*	*1*	*6*
	AUC	*2*	*3*	*4*	*5*	*1*	*6*
	Sensitivity	*2*	*3*	*5*	*4*	*1*	*6*
	Rank	***2 (2.00)***	***3 (3.00)***	***4 (4.33)***	***5 (4.67)***	***1 (1.00)***	***6 (6.00)***
ESIC	Accuracy	*6*	*5*	*2*	*4*	*1*	*3*
	AUC	*1*	*2*	*5*	*3*	*4*	*6*
	Sensitivity	*1*	*2*	*6*	*3*	*4*	*5*
	Rank	***1 (2.67)***	***2 (3.00)***	***5 (4.33)***	***4 (3.33)***	***2 (3.00)***	***6 (4.67)***
Rank for all data sets		***1 (2.00)***	***2 (2.56)***	***5 (4.44)***	***4 (4.11)***	***2 (2.56)***	***6 (5.33)***

Table 7: Rank per heterogenous data set and AF for the performance measures accuracy, sensitivity, and AUC. Mean values of the aggregated ranks are given in parentheses. ESIC: elective surgery and intensive care

Overall, the results of the sensitivity analysis show that the deep learning model involving our two AFs, *DENLU* and *leaky stanh,* is robust. The varying values of the parameters lead to varying outcomes at a low scale without significantly influencing the overall performance. For the parameters a and b,

it might be beneficial to use the existing values according to Liew et al. (2016) (i.e., $a = 1.7321$ and $b = 0.6585$), since the optimal values for these have already been sufficiently investigated. As the data structure differs among healthcare applications, we recommend adjusting parameter γ depending on the data set using hyperparameter tuning, despite undetectable differences in our data sets.

4.3.2 Comparison

DENLU and *leaky stanh* perform similarly to the other AFs in terms of accuracy for the heterogenous data sets (i.e., simulated, ESIC and Covid-19 triage data). The maximum deviation is -2.5 PP. Only the Covid-19 diagnosis data results in slightly worse performance in general, which can be attributed to the amount and the homogeneity of the data. In contrast to the other two data sets, the Covid-19 diagnosis data solely considers laboratory data. Since the data structure can significantly influence the performance of machine learning models, we recommend the application of our two AFs for heterogeneous data. In healthcare settings, sensitivity, and the integrated performance metric AUC (calculated from sensitivity and false positive rate) are of special interest, because the correct classification of patients in the positive class significantly contributes to optimal medical care and efficient and effective planning in hospitals. For the performance metrics, sensitivity, and AUC, *DENLU* and *leaky stanh* perform in most cases superior. We have identified a significant sensitivity gap for *DENLU* and *leaky stanh* compared with existing AFs. *Leaky stanh*'s performance is always close to the performance of *DENLU*. Figure *7* provides a comparison of the performance measures AUC and sensitivity for all test data sets and AFs.

When considering the rank for performance measures accuracy, sensitivity, and AUC across the heterogeneous data sets, *DENLU* achieved the highest rank, followed by *leaky stanh* and *sigmoid* (see Table *7*). The results demonstrate that decision-makers can receive improved assistance when making decisions within a specific class, such as ICU treatment, by incorporating AFs like *DENLU* or *leaky stanh*.

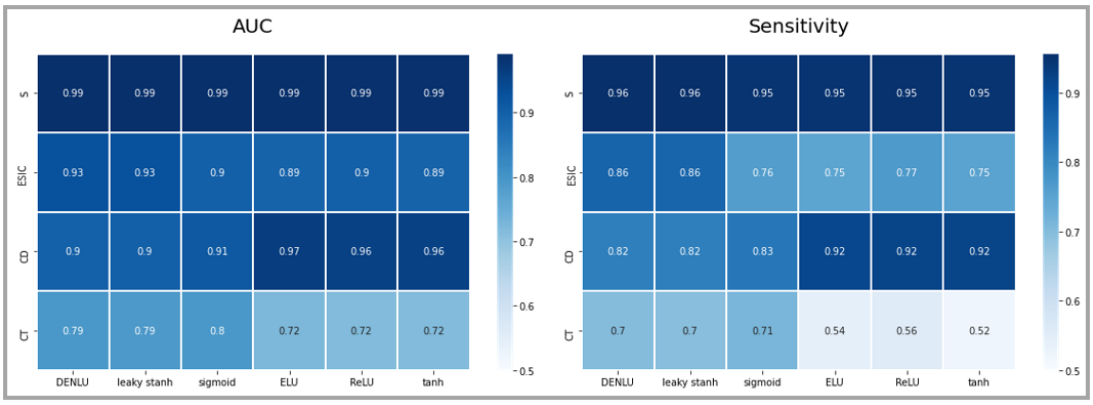

Figure 7: Comparison of performance measures AUC and sensitivity (average) for different AFs and test data sets (S: simulated, ESIC: elective surgery and intensive care, CD: Covid-19 diagnosis, CT: Covid-19 triage)

5 Conclusion

In this work, we present extended versions of existing AFs in machine learning with the objective of enhanced sensitivity: *DENLU* and *leaky stanh*. From a methodological point of view, *DENLU* is a flexible alternative to *sigmoid*, offering greater shaping capabilities. *Leaky stanh* is a sigmoidal piecewise linear alternative to *(s)tanh* that avoids the issue of vanishing gradients. *Sigmoid* and *tanh* have formed the basis for a large part of the AFs used in the field of healthcare and medicine for quite a long time. When applying our proposed AFs to simulated data and healthcare data for binary classification of patients, we have found that they perform better in terms of AUC (up to +7.6 PP) and especially sensitivity (up to +17.7 PP), guaranteeing their methodological advantages in our application domain. Precise predictions for a particular class, e.g., sensitivity, empower a decision-maker to create more precise plans, such as determining the need for ICU treatment.

Our computational analyses show some limitations which are discussed in the following. First, the data quality regarding missing or invalid entries varies among real-world data sets with corresponding influence on the outcome. As we have focused on the development and validation of extended AFs, we decided to set up a uniform procedure for data preparation rather than identifying the optimal data science procedure per data set. For example, there might exist superior procedures for feature scaling in the Covid-19 diagnosis

data leading to a higher level of performance across all AFs. Second, the structure of the neural network with a predefined number of layers and neurons might also influence the results but significantly contributes to comparability. In addition, our AFs provide better results solely for heterogeneous data sets, so more insight needs to be gained for homogeneous data.

The research field of developing methods for enhanced sensitivity is currently overshadowed by the focus on applying machine learning in digital healthcare. Future research should aim to bridge the gap by addressing both methodology and the practical application of these methods. An example is creating a decision support tool for selecting the appropriate AF for healthcare data sets.

Acknowledgements

For the acquisition of Covid-19 diagnosis and triage data, we express our deep gratitude to Dr. Lutz T. Zabel and Prof. Dr. Andreas Schuler of Alb-Fils-Kliniken Göppingen, Prof. Dr. Reinhard Hoffmann of the University Hospital of Augsburg, and all study teams supporting the LEOSS study. The LEOSS study group contributed at least 5 per mille to the analyses of this study: University Hospital Regensburg (Frank Hanses), Technical University of Munich (Christoph Spinner), Hospital Ingolstadt (Stefan Borgmann), University Hospital Freiburg (Siegbert Rieg), Klinikum Dortmund gGmbH, Hospital of University Witten / Herdecke (Martin Hower), University Hospital Frankfurt (Maria Vehreschild), University Hospital Jena (Maria Madeleine Ruethrich), Practice at Ebertplatz Cologne (Christoph Wyen), Hospital Bremen-Center (Bernd Hertenstein), Hospital Passau (Julia Lanznaster), University Hospital Augsburg (Christoph Roemmele), Johannes Wesling Hospital Minden Ruhr University Bochum (Kai Wille), Hospital Ernst von Bergmann (Lukas Tometten), University Hospital Essen (Sebastian Dolff), University Hospital Munich/ LMU (Michael von Bergwelt-Baildon), University Hospital Heidelberg (Uta Merle), Robert-Bosch-Hospital Stuttgart (Katja Rothfuss), University Hospital Wuerzburg (Nora Isberner), University Hospital Cologne (Norma Jung), University Hospital Tuebingen (Siri Göpel), Hospital Maria Hilf GmbH Moenchengladbach (Juergen vom Dahl), Municipal Hospital Karlsruhe (Christian Degenhardt), University Hospital Erlangen (Richard Strauss), University Hospital Ulm (Beate Gruener), Hospital Leverkusen (Lukas Eberwein), Catholic Hospital Bochum (St. Josef Hospital) Ruhr University Bochum (Kerstin Hellwig), Bundeswehr Hospital Koblenz (Dominic Rauschning), Evangelisches Hospital Saarbruecken (Mark Neufang), Marien Hospital Herne Ruhr University Bochum (Timm Westhoff), Tropical Clinic Paul-Lechler Hospital Tuebingen (Claudia Raichle), Hacettepe University (Murat Akova), University Hospital Duesseldorf (Bjoern-Erik Jensen), Elbland Hospital Riesa (Joerg Schubert), Center for Infectiology Prenzlauer Berg Berlin (Stephan Grunwald), University Hospital Schleswig-Holstein Kiel (Anette Friedrichs), University Hospital of Giessen and Marburg (Janina Trauth), University Hospital Dresden (Katja de With), Clinic Munich (Wolfgang Guggemos), Hospital Braunschweig (Jan Kielstein), Agaplesion Diakonie Hospital Rotenburg (David Heigener), Hospital Fulda (Philipp Markart), University Hospital Saarland (Robert Bals), Petrus Hospital Wuppertal (Sven Stieglitz), Elisabeth Hospital Essen (Ingo Voigt), Richmond Research Institute (Jorg Taubel), Malteser Hospital St. Franziskus Flensburg (Milena Milovanovic). The LEOSS study infrastructure group: Jörg Janne Vehreschild (Goethe University Frankfurt), Susana M. Nunes de Miranda (University Hospital of Cologne), Carolin E. M. Jakob (University Hospital of Cologne), Melanie Stecher (University Hospital of Cologne), Lisa Pilgram (Goethe University Frankfurt), Nick Schulze (University Hospital of Cologne), Sandra Fuhrmann (University Hospital of Cologne), Max Schons (University Hospital of Cologne), Annika Claßen (University Hospital of Cologne), Bernd Franke (University Hospital of Cologne) und Fabian Praßer (Charité, Universitätsmedizin Berlin).

For the acquisition of elective surgery and intensive care data, we express our deep gratitude to Dr. Thomas Koperna, former operating room manager at University Hospital of Augsburg.

Financial support

No specific financial support was received for this study. The LEOSS registry was supported by the German Centre for Infection Research (DZIF) and the Willy Robert Pitzer Foundation.

Conflicts of interest

The authors declare no conflicts of interest.

References

Alballa N, Al-Turaiki I (2021) Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Informatics in Medicine Unlocked 24:100564. https://doi.org/10.1016/j.imu.2021.100564

Chandra P, Ghose U, Sood A (2015) A non-sigmoidal activation function for feedforward artificial neural networks. In: 2015 International Joint Conference on Neural Networks (IJCNN). IEEE

Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Diseases 40:373–383. https://doi.org/10.1016/0021-9681(87)90171-8

Chung H, Lee SJ, Park JG (2016) Deep neural network using trainable activation functions. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE

Das D, Pasupathy KS, Storlie CB, Sir MY (2019) Functional regression-based monitoring of quality of service in hospital emergency departments. IISE Transactions 51:1012–1024. https://doi.org/10.1080/24725854.2018.1536303

El-Baz A, Suri JS (2021) Machine learning in medicine. Chapman & Hall/CRC healthcare informatics series. CRC Press, Taylor and Francis Group, Boca Raton

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Networks 4:251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Li J-C, Ng WWY, Yeung DS, Chan PPK (2014) Bi-firing deep neural networks. Int. J. Mach. Learn. & Cyber. 5:73–83. https://doi.org/10.1007/s13042-013-0198-9

Liew SS, Khalil-Hani M, Bakhteri R (2016) Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. Neurocomputing 216:718–734. https://doi.org/10.1016/j.neucom.2016.08.037

Ohn I, Kim Y (2019) Smooth Function Approximation by Deep Neural Networks with General Activation Functions. Entropy 21:627. https://doi.org/10.3390/e21070627

Pfeuffer N, Baum L, Stammer W, Abdel-Karim BM, Schramowski P, Bucher AM, Hügel C, Rohde G, Kersting K, Hinz O (2023) Explanatory Interactive Machine Learning. Bus Inf Syst Eng:1–25. https://doi.org/10.1007/s12599-023-00806-x

Puheim M, Nyulaszi L, Madarasz L, Gaspar V (2014) On practical constraints of approximation using neural networks on current digital computers. In: IEEE 18th International Conference on Intelligent Engineering Systems INES 2014. IEEE

Sodhi SS, Chandra P (2014) Bi-modal derivative activation function for sigmoidal feedforward networks. Neurocomputing 143:182–196. https://doi.org/10.1016/j.neucom.2014.06.007

Weber M, Beutter M, Weking J, Böhm M, Krcmar H (2022) AI Startup Business Models. Bus Inf Syst Eng 64:91–109. https://doi.org/10.1007/s12599-021-00732-w

Wynants L, van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Dahly DL, Damen JAA, Debray TPA, Jong VMT de, Vos M de, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Heus P, Kammer M, Kreuzberger N, Lohmann A, Luijken K, Ma J, Martin GP, McLernon DJ, Andaur Navarro CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM,

Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, Tzoulaki I, van Kuijk SMJ, van Bussel B, van der Horst ICC, van Royen FS, Verbakel JY, Wallisch C, Wilkinson J, Wolff R, Hooft L, Moons KGM, van Smeden M (2020) Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369:m1328. https://doi.org/10.1136/bmj.m1328

Appendix

Tables

Measure	Data set	***DENLU***	***leaky stanh***	*ELU*	*ReLU*	*sigmoid*	*tanh*
Accuracy	Simulated	**0.9560**	0.9559	0.9521	0.9510	0.9536	0.9494
	Covid-19 triage	0.7265	0.7240	0.7194	0.7172	**0.7324**	0.7115
	Covid-19 diagnosis	0.8226	0.8234	0.9098	**0.9114**	0.8351	0.9074
	ESIC	0.8670	0.8679	0.8883	0.8834	**0.8916**	0.8878
AUC	Simulated	**0.9915**	0.9914	0.9899	0.9894	0.9914	0.9881
	Covid-19 triage	0.7949	0.7905	0.7228	0.7187	**0.7953**	0.7186
	Covid-19 diagnosis	0.8982	0.8997	**0.9653**	0.9592	0.9132	0.9585
	ESIC	**0.9294**	0.9287	0.8945	0.9007	0.8995	0.8940
F1	Simulated	**0.9560**	**0.9560**	0.9519	0.9509	0.9533	0.9494
	Covid-19 triage	0.5815	0.5774	0.5112	0.5155	**0.5901**	0.4944
	Covid-19 diagnosis	0.8678	0.8686	0.9357	**0.9370**	0.8783	0.9342
	ESIC	0.5748	0.5799	0.5813	0.5785	**0.5942**	0.5829
NPV	Simulated	0.9562	**0.9563**	0.9480	0.9518	0.9490	0.9488
	Covid-19 triage	0.8677	0.8655	0.8216	0.8245	**0.8736**	0.8153
	Covid-19 diagnosis	0.6462	0.6467	0.8094	**0.8188**	0.6643	0.8168
	ESIC	**0.9822**	0.9820	0.9709	0.9709	0.9705	0.9691
Precision	Simulated	0.9559	0.9555	0.9562	0.9502	**0.9584**	0.9501
	Covid-19 triage	0.4991	0.4949	0.4853	0.4850	**0.5048**	0.4719
	Covid-19 diagnosis	0.9253	0.9265	**0.9559**	0.9532	0.9311	0.9475
	ESIC	0.4309	0.4371	0.4765	0.4647	**0.4870**	0.4761
Sensitivity	Simulated	0.9561	**0.9564**	0.9477	0.9517	0.9484	0.9488
	Covid-19 triage	0.6994	0.6951	0.5417	0.5563	**0.7118**	0.5226
	Covid-19 diagnosis	0.8174	0.8180	0.9169	**0.9215**	0.8314	0.9215
	ESIC	**0.8643**	0.8633	0.7466	0.7629	0.7529	0.7523

Table 8: Comparison of performance measures (average) for different AFs and test data sets

Data set	**Input layer**	**Hidden layer 1**	**Hidden layer 2**	**Hidden layer 3**	**Hidden layer 4**	**Output layer**	**Learning rate**
S	N: 460	N: 460	N: 10			AF: *sigmoid*	0.006
CT	N: 160	N:310				AF: *sigmoid*	0.002
CD	N: 320	N: 10	N: 10	N: 10	N: 10	AF: *softmax*	0.002
ESIC	N: 110	N: 10	N: 10	N: 10		AF: *sigmoid*	0.006

Table 9: Changes in the neural network compared to the original neural network for each data set after performing hyperparameter tuning for the AF DENLU (S: simulated, CT: Covid-19 triage, CD: Covid-19 diagnosis, ESIC: elective surgery and intensive care, N: neurons, AF: activation function)

Additional material: Differences between *DENLU* and *stanh*

$$\sigma_{tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \text{ and } \sigma_{DENLU}(x) = \begin{cases} e^x - 1\,, & \text{if } x \leq 0 \\ 1 - e^{-x}, & \text{if } x > 0 \end{cases}$$

i) $x > 0$:

$$\sigma_{DENLU}(x) = 1 - e^{-x} = \frac{e^x + e^{-x}}{e^x + e^{-x}} - e^{-x} = \frac{e^x + e^{-x}}{e^x + e^{-x}} - \frac{e^{-x} \cdot (e^x + e^{-x})}{e^x + e^{-x}}$$

$$\sigma_{DENLU}(x) = \frac{e^x + e^{-x} - e^{-x} \cdot e^x - e^{-x} \cdot e^{-x}}{e^x + e^{-x}} = \frac{e^x + e^{-x} - 1 - e^{-2x}}{e^x + e^{-x}}$$

$$\sigma_{DENLU}(x) - \sigma_{tanh}(x) = \frac{e^x + e^{-x} - 1 - e^{-2x}}{e^x + e^{-x}} - \frac{e^x - e^{-x}}{e^x + e^{-x}}$$
$$= \frac{e^x + e^{-x} - 1 - e^{-2x} - e^x + e^{-x}}{e^x + e^{-x}}$$

$$\sigma_{DENLU}(x) - \sigma_{tanh}(x) = -\frac{1 - 2e^{-x} + e^{-2x}}{e^x + e^{-x}} = -\frac{(1 - e^{-x})^2}{e^x + e^{-x}}$$

ii) $x \leq 0$:

$$\sigma_{DENLU}(x) = e^x - 1 = \frac{e^x \cdot (e^x + e^{-x})}{e^x + e^{-x}} - \frac{e^x + e^{-x}}{e^x + e^{-x}} = \frac{e^x \cdot (e^x + e^{-x}) - e^x - e^{-x}}{e^x + e^{-x}}$$

$$\sigma_{DENLU}(x) = \frac{e^{2x} + e^x \cdot e^{-x} - e^x - e^{-x}}{e^x + e^{-x}} = \frac{-e^x - e^{-x} + 1 + e^{2x}}{e^x + e^{-x}}$$

$$\sigma_{DENLU}(x) - \sigma_{tanh}(x) = \frac{-e^x - e^{-x} + 1 + e^{2x}}{e^x + e^{-x}} - \frac{e^x - e^{-x}}{e^x + e^{-x}}$$
$$= \frac{-e^x - e^{-x} + 1 + e^{2x} - e^x - e^{-x}}{e^x + e^{-x}}$$

$$\sigma_{DENLU}(x) - \sigma_{tanh}(x) = \frac{1 - 2e^x + e^{2x}}{e^x + e^{-x}} = \frac{(1 - e^x)^2}{e^x + e^{-x}}$$

Figures

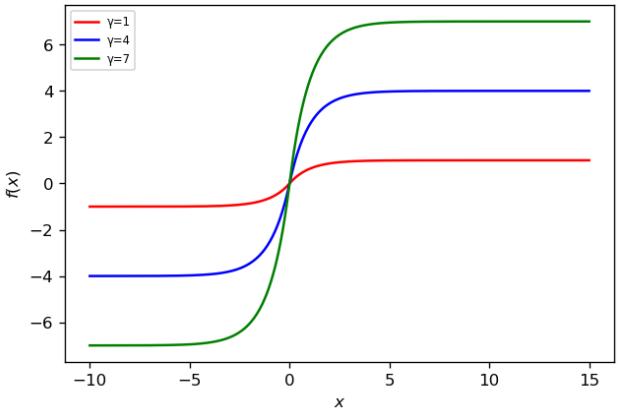

Figure 8: DENLU for different values of γ

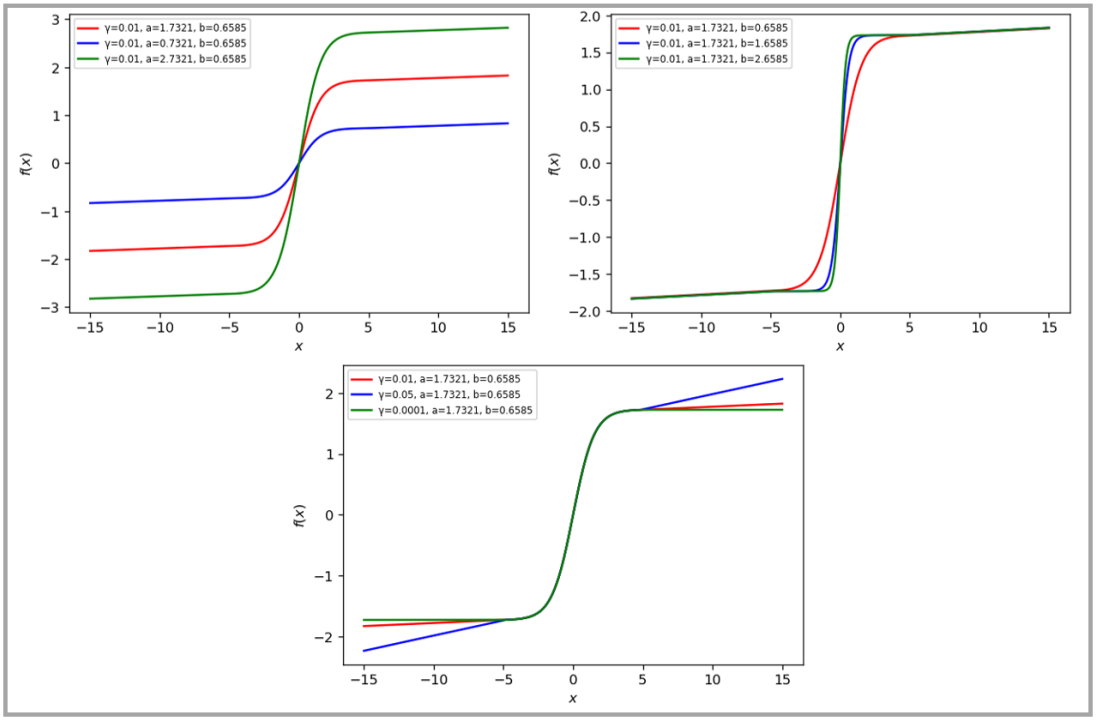

Figure 9: Leaky stanh for different values of a, b and γ

Appendix C: Covid-19 triage

Bartenschlager, CC, Grieger, M, Erber, J, Neidel, T, Borgmann, S, Vehreschild, JJ, Steinbrecher, M, Rieg, S, Stecher, M, Dhillon, C, Ruethrich, MM, Jakob, CEM, Hower, M, Heller, AR, Vehreschild, M, Wyen, C, Messmann, H, Pipel, C, Brunner, JO, Hanses, F, Römmele, C (2023). Covid-19 triage in the emergency department 2.0: How analytics and AI transform a human-made algorithm for the prediction of clinical pathways.

Status: Published in *Health Care Management Science*; Category A.

The final version of the contribution is available online at https://doi.org/10.1007/s10729-023-09647-2.