

Bohlen, Lasse; Rosenberger, Julian; Zschech, Patrick; Kraus, Mathias

Article — Published Version

Leveraging interpretable machine learning in intensive care

Annals of Operations Research

Provided in Cooperation with:

Springer Nature

Suggested Citation: Bohlen, Lasse; Rosenberger, Julian; Zschech, Patrick; Kraus, Mathias (2024) : Leveraging interpretable machine learning in intensive care, Annals of Operations Research, ISSN 1572-9338, Springer US, New York, NY, Vol. 347, Iss. 2, pp. 1093-1132, <https://doi.org/10.1007/s10479-024-06226-8>

This Version is available at:

<https://hdl.handle.net/10419/323300>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<http://creativecommons.org/licenses/by/4.0/>



Leveraging interpretable machine learning in intensive care

Lasse Bohlen¹  · Julian Rosenberger² · Patrick Zschech³  · Mathias Kraus² 

Received: 31 March 2023 / Accepted: 15 August 2024 / Published online: 19 September 2024
© The Author(s) 2024

Abstract

In healthcare, especially within intensive care units (ICU), informed decision-making by medical professionals is crucial due to the complexity of medical data. Healthcare analytics seeks to support these decisions by generating accurate predictions through advanced machine learning (ML) models, such as boosted decision trees and random forests. While these models frequently exhibit accurate predictions across various medical tasks, they often lack interpretability. To address this challenge, researchers have developed interpretable ML models that balance accuracy and interpretability. In this study, we evaluate the performance gap between interpretable and black-box models in two healthcare prediction tasks, mortality and length-of-stay prediction in ICU settings. We focus specifically on the family of generalized additive models (GAMs) as powerful interpretable ML models. Our assessment uses the publicly available Medical Information Mart for Intensive Care dataset, and we analyze the models based on (i) predictive performance, (ii) the influence of compact feature sets (i.e., only few features) on predictive performance, and (iii) interpretability and consistency with medical knowledge. Our results show that interpretable models achieve competitive performance, with a minor decrease of 0.2–0.9 percentage points in area under the receiver operating characteristic relative to state-of-the-art black-box models, while preserving complete interpretability. This remains true even for parsimonious models that use only 2.2% of patient features. Our study highlights the potential of interpretable models to improve decision-making in ICUs by providing medical professionals with easily understandable and verifiable predictions.

Keywords Healthcare analytics · Interpretable machine learning · Generalized additive models · Length-of-stay prediction · Mortality prediction

Lasse Bohlen and Julian Rosenberger have contributed equally to this work.

✉ Lasse Bohlen
lasse.bohlen@fau.de

Julian Rosenberger
julian.rosenberger@ur.de

Patrick Zschech
patrick.zschech@uni-leipzig.de

Mathias Kraus
mathias.kraus@ur.de

¹ Friedrich-Alexander-Universität Erlangen-Nürnberg, Lange Gasse 20, 90403 Nürnberg, Germany

² Universität Regensburg, Bajuwarenstraße 4, 93053 Regensburg, Germany

³ Universität Leipzig, Grimmaische Str. 12, 04109 Leipzig, Germany

1 Introduction

Healthcare systems across the globe face a multitude of complex challenges, such as care quality disparities, demographic shifts, and administrative obstacles including resource shortages, rising costs, and insufficient infrastructure (Roncarolo et al., 2017). These issues are intensified by increasing demand for healthcare services, sophisticated medical technology complicating physicians' workflows, and heightened expectations for patient-centered care. In the context of intensive care units (ICUs), these challenges are further heightened due to patient acuity, the need for specialized staff, and pressure on resource allocation. ICUs account for approximately 14 % of total hospital expenses, making them one of the most critical and costly components of healthcare systems (Halpern & Pastores, 2010). Consequently, effective ICU management is essential for optimizing patient outcomes and ensuring healthcare systems' sustainability (Bertsimas et al., 2021). However, optimizing resource utilization in ICUs remains a daunting task due to the urgent and unpredictable nature of intensive care, high costs associated with maintaining sufficient resources, and constant demand for specialized staff (Bai et al., 2018).

One promising strategy to address ICU challenges is the implementation of machine learning (ML) models. ML models, with their ability to make predictions more quickly and on a larger scale than humans, are increasingly considered indispensable for modern healthcare systems (Malik et al., 2018). Yet, advanced ML models such as boosted decision trees and random forests are often regarded as the primary ML models with superior predictive performance (e.g., Hyland et al., 2020), despite their black-box characteristics, which render their decision logic difficult for humans to understand. This perception has led to a widespread belief that interpretability must be compromised to achieve accurate predictions (Gunning & Aha, 2019), significantly impeding ML adoption in healthcare (Kundu, 2021). While this viewpoint has been challenged by numerous researchers (e.g., Caruana et al., 2015; Rudin, 2019), the current literature lacks an in-depth analysis examining the performance gap between black-box and interpretable models.

A fundamental principle of achieving interpretable ML models is to use the simplest possible model that adequately explains the data (Brunton & Kutz, 2019). This concept is also known as model parsimony. In general, there are two objectives to focus on in order to obtain a parsimonious model. First, the ML model itself should be interpretable, i.e., of limited complexity, so that humans can understand the model, and second, the number of features used by the model to compute an output should be small, i.e., using a compact feature set.

Our study seeks to investigate the performance differences along these two objectives. We compare black-box and interpretable models, as well as the role of compact feature sets. By illustrating that interpretable models with a minimal set of input features can attain comparable accuracy, we aim to promote further research and development of interpretable ML models for healthcare applications. This work contributes to a more comprehensive understanding of interpretability in ML models and emphasizes the significance of these factors for successful implementation in healthcare settings. Specifically, this work contributes to the existing literature in four ways:

1. We examine and evaluate multiple interpretable models from the family of generalized additive models (GAMs) against prevalent black-box models in an ICU setting. We com-

pare three distinct prediction targets: length-of-stay > 3 days (LOS3), length-of-stay > 7 days (LOS7), and mortality.¹

2. We perform a comparative analysis to evaluate the impact of feature reduction methods on predictive performance, with particular emphasis on sensitive features such as gender, age, and ethnicity.
3. We showcase the utility of interpretable models by discussing their plausibility with four medical experts from diverse backgrounds.
4. To make these models competitive, we also propose multiple feature engineering steps that yield favorable results in comparison to previous approaches.²

Our findings challenge the prevailing belief that only black-box models provide high predictive performance in healthcare. We demonstrate that interpretable models can achieve competitive predictive performance, with a minor decrease of 0.2–0.9 percentage points in area under the receiver operating characteristic (AUROC) compared to black-box models, while remaining full interpretability. This finding holds true even for parsimonious models that utilize only 2.2 % of the patient features available, while exhibiting a negligible performance drop relative to black-box models, ranging from 0.1 to 1.0 percentage points and averaging at 0.5 percentage points. By showcasing the comparable accuracy of interpretable models even with compact feature sets, we aim to inspire further research and development of interpretable ML models in healthcare applications.

The remainder of the paper is structured as follows: Sect. 2 delves into the conceptual background and prior research on ML in healthcare, legal and practical requirements for ML models, and interpretable ML. Section 3 details the dataset, prediction tasks, used models, and the experiments. In Sect. 4, we evaluate and compare the proposed models against black-box models, and visually inspect so-called shape plots produced by GAMs. Section 5 discusses the implications of our work for research and practice, along with its limitations. Section 6 concludes the paper.

2 Research background

2.1 Generalized additive models

This study emphasizes the use of GAMs as a particular powerful class of interpretable models. GAMs have been employed in other high-stakes domains where model interpretability is essential (Chang et al., 2021; Zschech et al., 2022). Fundamentally, GAMs are ML models that build relationships between input features and the target by summing several distinct univariate non-linear mappings, called shape functions. Formally, GAMs can be expressed as

$$f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m), \quad (1)$$

where each f_i denotes a shape function mapping the i -th input feature to the output space. As such, GAMs preclude feature-interactions, potentially sacrificing model performance but enabling complete comprehension of model functionality. Building on this core concept of utilizing nonlinear functions to map input features to the output space, various GAM versions

¹ Our evaluation pipeline and replication code can be found here:
https://github.com/HBDynamite/Interpretable_ICU_predictions

² For our feature engineering steps, see:
https://github.com/HB-Dynamite/mimic3-benchmarks_AoOR_data_export

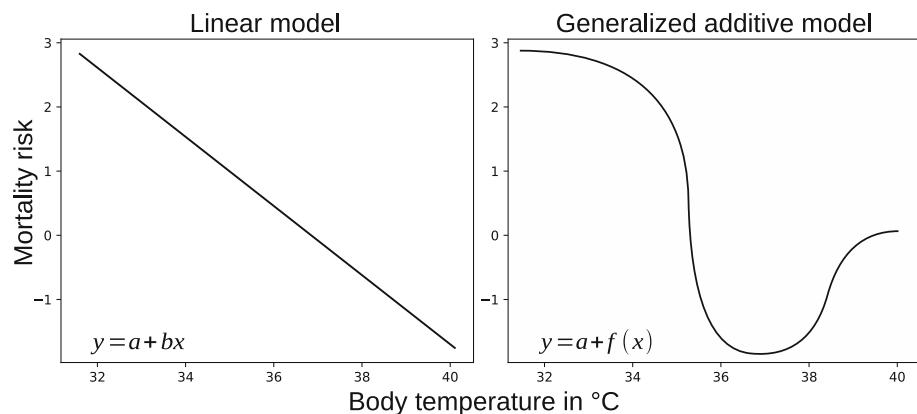


Fig. 1 Comparison of the shape functions of a linear model and a GAM

have been proposed (Lou et al., 2012; Agarwal et al., 2021; Kraus et al., 2024b). A GAM can also function as a classification model by modeling the log odds of the target class probabilities. The log odds are the logarithm of the odds, which are the ratio of the target class probabilities. In a binary classification context, the GAM can be expressed as

$$\log \left(\frac{P(y = 1)}{1 - P(y = 1)} \right) = f_1(x_1) + f_2(x_2) + \dots + f_m(x_m), \quad (2)$$

where $P(y = 1)$ denotes the probability of the target class given features x_1, \dots, x_m , and $f_i(x_i)$ are shape functions. The logit function maps the probability space $[0, 1]$ onto $(-\infty, \infty)$, allowing the model to predict target class probabilities. To obtain class probabilities, the logistic function is applied to the model output.

The independence of the input features enables a 2-dimensional visualization of shape functions using so-called shape plots. Figure 1 presents an example for a shape plot, where the GAM captures the nonlinear relationship between body temperature and the probability (in log odds) of mortality (represented by the solid curve). In contrast, the linear model assumes a linear relationship (represented by the straight line). This comparison highlights the flexibility of GAMs in modeling complex patterns, providing more accurate and interpretable predictions than linear alternatives. For physiological signs, often an optimal value exists, with deviations in either direction increasing mortality. Eventually, human experts can evaluate the established relationships through this visualization (Hegselmann et al., 2020).

Furthermore, GAMs have the advantage of being intrinsically interpretable (Du et al., 2019). That is, they provide an exact description of their decision logic rather than an approximation. This is in stark contrast to post-hoc explanation methods such as Shapley additive explanations (SHAP), where complex ML models are simplified based on rough approximations to gain certain insights into the models' behavior (Stiglic et al., 2020; Rudin, 2019). Such post-hoc explanations inevitably lead to a loss of information, which creates the risk of erroneous insights and thus a harmful basis for medical decision support (Babic et al., 2021). Therefore, intrinsically interpretable models such as GAMs offer a more reliable choice because they provide an undistorted view of the global model structure, which can be easily analyzed, adjusted, and validated for safety and efficacy.

2.2 Healthcare decisions and machine learning

Advanced ML models, such as boosted decision trees and random forests, have transformed healthcare analytics by enabling extensive medical data analysis (Bohr & Memarzadeh, 2020; Saadatmand et al., 2022; Hyland et al., 2020). However, their complex structure and black-box behavior have slowed their adoption in high-stakes domains (Miller, 2019; Rudin, 2019). The models' lack of interpretability makes it difficult to understand what drives their predictions. This lack of clarity fosters skepticism among regulators and practitioners, hindering the widespread use of these powerful ML techniques (Agarwal & Das, 2020).

Despite skepticism, ML models in healthcare provide numerous benefits, such as surpassing human performance in diverse clinical decision support areas (Richens et al., 2020). Human judgment is limited by perceptual biases and cognitive constraints, while machines can process data more quickly and efficiently. Some algorithmic architectures even surpass doctors in predictive accuracy (Johnson et al., 2022). ML models can also aid essential resource allocation, such as bed allocation or work scheduling, by providing decision-makers with real-time information about the entire patient population (Bertsimas et al., 2021). This approach aligns with operations research principles, as ML models can be integrated with optimization techniques to improve healthcare decision-making (Bai et al., 2018; Johnson et al., 2022; Kraus et al., 2024a). However, it is crucial to recognize the challenges and limitations associated with implementing ML models in healthcare settings. These encompass legal, practical, and ethical issues, which we discuss in the following section.

2.3 Legal, practical, and ethical requirements in healthcare

The legal landscape for algorithmic interpretability is evolving, with authorities utilizing primarily recommendations, guidelines, and preliminary frameworks. In the United States, an exemption for interpretable medical software has been introduced (Gerke et al., 2020) and is overseen by the Food and Drug Administration (FDA), which is responsible for regulating medical devices. In the European Union (EU), the focus is on promoting algorithmic transparency. The EU enforces a "right to explanation" under the General Data Protection Regulation (GDPR) (Parliament and Council of the European Union, 2016; Goodman & Flaxman, 2017). This mandate emphasizes the importance of interpretability for protecting sensitive data³ and ensuring fair and ethical treatment. Regulatory efforts may become more stringent, as the EU is developing the Artificial Intelligence Act (Parliament and Council of the European Union, 2021), which aims to regulate ML use in high-stakes decision-making. This legislation could further emphasize the significance of algorithmic interpretability in the U.S. and EU, promoting interpretable and accountable ML systems to ensure ethical and legal compliance.

From a practical perspective, healthcare decision-makers prioritize patient well-being above all else, and typically do not possess extensive knowledge of ML techniques. Therefore, ML model interpretability is essential to enable care givers to perform informed decision-making rather than blindly relying on opaque predictions (Stiglic et al., 2020). Explanations should align with user skills and domain knowledge, reducing the risk of application errors (Coussement & Benoit, 2021).

³ "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs; trade-union membership; genetic data, biometric data processed solely to identify a human being; health-related data; data concerning a person's sex life or sexual orientation" [Article 4 (13), (14) and (15) and Article 9 and Recitals (51) to (56) of the GDPR].

Regarding ethical requirements, handling sensitive data demands special attention to protect individual privacy and prevent discriminatory practices (Goodman & Flaxman, 2017). The potentially life-altering consequences of healthcare amplify these concerns (Babic et al., 2021). Inherently interpretable models, such as GAMs, address ethical concerns related to sensitive data by providing insights into the impact of individual features on predictions (Chang et al., 2021). This interpretability facilitates the identification and mitigation of potential biases, ensuring fair and accurate healthcare decisions.

2.4 Achieving model interpretability

Model interpretability refers to a model's ability to present its behavior in human-understandable terms (Doshi-Velez & Kim, 2017). The necessary level of interpretability depends on the use case and domain, as stakeholders may have varying expectations and requirements. What is sufficient for one use case may not be for another (Rudin, 2019).

In healthcare, risk charts such as the Framingham Risk Score for predicting the 10-year risk of developing cardiovascular diseases (Wilson et al., 1998), the Simplified Acute Physiology Score (SAPS), and Acute Physiology and Chronic Health Evaluation (APACHE) scores for severity-of-illness classification in ICUs (Moreno et al., 2005) are commonly used to assess patient prognosis and inform clinical decision-making. These risk charts represent simple, interpretable models that enable clinicians to quickly understand patient conditions and make informed decisions.

Interpretability and ease of understanding are crucial factors when selecting ML models in the healthcare domain. While more complex ML models can often provide more accurate assessments of patients' conditions, simpler models such as decision trees and linear models have been favored for their interpretability. Decision trees provide visual representations that allow physicians to trace decision-making processes (Breiman et al., 1984), making it easier to understand how the model arrived at its predictions. Similarly, linear models help understand the contribution of each feature to the predictions, providing insights into the factors that influence the model's output. By maintaining interpretability, these techniques foster trust and adoption in healthcare settings, where understanding the reasoning behind predictions is essential for making informed decisions (Kundu, 2021).

GAMs have emerged as powerful tools that strike a balance between accuracy and interpretability, making them well-suited for healthcare applications (Zschech et al., 2022). GAMs combine the simplicity of linear models with the flexibility of nonlinear functions, capturing complex relationships between input features and outcomes while preserving interpretability (Chang et al., 2021). The renewed interest in GAMs has been fueled by integrating advanced techniques like boosting and specifically designed neural networks (Yang et al., 2021; Lou et al., 2012; Agarwal et al., 2021).

Model parsimony, which refers to the ability of a model to describe the data using the minimum number of terms or parameters, is closely related to interpretability. Parsimonious models strike a balance between fitting the data well and avoiding unnecessary complexity (Brunton & Kutz, 2019). Thus, parsimony promotes simplicity while retaining a high level of accuracy. This is achieved through two objectives. First, the model architecture should be kept as simple as necessary, which also allows people to understand its functionality. Second, the model should be based only on a subset of selected features, identifying the most informative ones and reducing the dimensionality of the input space in order to obtain a compact model (James et al., 2013). GAMs in particular promote parsimony by their architecture, as they are typically simpler than more complex models such as neural networks. By combining

GAMs with feature selection techniques, we can ensure the development of parsimonious models that are both accurate and interpretable, similar to the risk maps commonly used in healthcare.

3 Research approach

This study aims to evaluate the performance gap between black-box and interpretable GAMs in an ICU setting. We use a well-established intensive care database and focus on two common binary classification tasks: mortality and length-of-stay prediction. The methodology is detailed in the subsequent sections, covering the dataset, prediction tasks, feature extraction, preprocessing steps, models, and analyses performed.

3.1 MIMIC-III clinical database

With the increasing adoption of healthcare information systems, hospitals start to generate and store large amounts of data. In particular, electronic health records track the patients' health trajectories combining information about demographics, physiological signs, and laboratory results. This data potentially allows to derive informed predictions about the patients' health (Kocheturov et al., 2019).

In this study, we use the Medical Information Mart for Intensive Care (MIMIC)-III database (v1.4), one of the most extensive healthcare database that is publicly available (Johnson et al., 2016). MIMIC-III contains 58,976 anonymized health records of 46,520 patients admitted to the ICU of the Beth Israel Deaconess Medical Center between 2001 and 2012.

3.2 Mortality and length-of-stay prediction

We evaluate our models on two common healthcare prediction tasks: in-hospital-mortality⁴ and length-of-stay. Mortality and length-of-stay are the primary cost drivers in ICUs (Kramer et al., 2017) and as Bates et al. (2014) emphasize, early identification of high-risk and high-cost patients is key for implementing strategies to conserve resources in care. Consequently, these high-stakes prediction tasks are well suited to study the performance gap between interpretable and black-box ML models.

Mortality. This binary classification task predicts patient survival or death based on physiological data collected during the first few hours after admission to the ICU (Harutyunyan et al., 2019; Wang et al., 2020). The goal is to identify high-risk patients for timely provision of intensified medical supervision, care, and resource allocation.

Length-of-stay. This task estimates which patients will have longer ICU stays and require additional resources. Even rough length-of-stay estimates can significantly improve ICU scheduling, as long-term patients account for a disproportionate share of ICU resources (Halpern & Pastores, 2010; Kramer et al., 2017). The implementation details of this prediction task vary across the literature, e.g., it is framed as a multiclass classification (Harutyunyan et al., 2019) or regression problem (Purushotham et al., 2018). For improved comprehensibility, we follow Wang et al. (2020) and divide the task into two binary classification tasks: length-

⁴ We predict in-hospital-mortality by determining whether patients will die during their hospital stay or survive to discharge. For readability, we use "mortality" as a synonym for in-hospital-mortality in this paper.

Table 1 Descriptive statistics of the target features: mortality, Length-of-stay3, Length-of-stay7 - stratified by gender

	Gender		Total
	Female	Male	
<i>Mortality</i>			
Alive	13,724	17,764	31,488
Dead	1701	1939	3640
<i>LOS</i>			
<3	9394	12,179	21,573
>3	6031	7524	13,555
<7	13,376	17,087	30,463
>7	2049	2616	4665
Total	15,425	19,703	35,128

of-stay > 3 days (LOS3) and length-of-stay > 7 days (LOS7), providing better performance comparability with mortality classification. The same feature sets are used to predict length-of-stay and mortality, potentially enabling decision-makers to gain insight into expected expenditures at the patient, ward, and hospital levels (Halpern & Pastores, 2010).

Table 1 presents descriptive statistics for the target features, mortality, LOS3, and LOS7, showing a significant imbalance in their distribution. For instance, mortality has a larger number of alive patients (31,488 [89.6 %]) compared to deceased patients (3640 [10.4 %]). Similarly, most patients have a length-of-stay below the respective thresholds for LOS3 and LOS7.

3.3 Feature extraction, preprocessing, and evaluation strategy

In this section, we describe the process of feature extraction, preprocessing, and the evaluation strategy for our ML models. Our goal is to create a clean and consistent database for analysis and comparison of the performance gap between interpretable and black-box models.

Feature extraction. Regarding feature extraction, we prioritize reproducibility by primarily relying on the widely-used MIMIC-III benchmark suite created by Harutyunyan et al. (2019). To create a patient-specific time series of physiological measurements, we utilize the corresponding scripts with the following modifications to the original approach:

- As suggested by Wang et al. (2020) and Purushotham et al. (2018), we reduce the considered time period from the first 48 h to the first 24 h of ICU stay.
- We remove implausible values from the time series, following the approach of Hegselmann et al. (2020), detailed in Appendix A.
- The often missing total score of the Glasgow Coma Scale is recalculated as the sum of the sub scores (Teasdale & Jennett, 1974) whenever possible, and subsequently the sub scores are removed from further evaluations.
- Three sensitive features are intentionally (see Sect. 2.3) included to explore their influence in detail: gender, age, and ethnicity.

After these steps, we obtain a cohort of 35,128 patients, displayed in Table 2 with a particular focus on patients' sensitive features.

To extract meaningful static features from these time series, we compute six sample statistics (mean, standard deviation, minimum, maximum, skewness, and number of measurements) for seven different subperiods (full time series and subsets representing the first

Table 2 Default cohort summary stratified by sensitive features adopted from Wang et al. (2020)

	Gender		Total
	Female	Male	
<i>Ethnicity</i>			
Asian	363	479	842
Hispanic	465	740	1205
Black	1892	1470	3362
Other	1816	2745	4561
White	10,889	14,269	25,158
<i>Age</i>			
< 30	690	909	1599
31–50	2303	3282	5585
51–70	5224	8266	13,490
> 70	7208	7246	14,454

and last 10%, 25%, and 50% of time), resulting in 42 (6×7) features per time series variable (Harutyunyan et al., 2019).

Preprocessing. We apply standard preprocessing steps to the extracted dataset to make it suitable for comparison, including standardization and removal of features with more than 50% missing values. Missing values are imputed using the mean. For the latter, we also experiment with more advanced approaches such as k-nearest neighbor imputation and iterative random forest (Stekhoven & Bühlmann, 2012), but with limited success. See Appendix B for results of these experiments.

Excluding the sensitive features gender, age, and ethnicity, our approach results in 11 features as shown in Table 3. These physiological features are clinical indicators and cover various medical examination areas, including basic life-sustaining functions like respiratory rate, cardiovascular measures like diastolic pressure, biological markers like urinary glucose, and awareness-related measurements like the Glasgow Coma Scale (GCS) (Johnson et al., 2016). They aim to provide a comprehensive description of patients' conditions within the first 24 h of ICU stays (Harutyunyan et al., 2019). Applying the feature generation described in the preceding paragraph results in our initial feature set, consisting of 462 features, excluding the three sensitive features: gender, age, and ethnicity.

Evaluation. In our evaluation strategy, we employ 5-fold cross-validation for assessing the models' performance, i.e., we split the dataset into five equal folds and use four folds for training (80%) and one fold for testing (20%). We repeat this process until each fold has been used as the test set once. This approach offers a more reliable estimate by mitigating the effects of random variations in the data. By measuring out-of-sample performance in each fold and calculating the mean and standard deviation of the model performance across all folds, we effectively compare the models.

In this study, we tackle the challenge of highly imbalanced targets (see Table 1) by using both the area under receiver operating characteristic (AUROC) and the area under precision recall curve (AUPRC) to evaluate our ML models' classification performance. The AUROC is a widely-used metric to measure classifier performance. However, in the context of imbalanced datasets, the AUROC may convey an overly optimistic view of the model's performance. To address this limitation, we also employ the AUPRC, which emphasizes the identification of the minority class (Davis & Goadrich, 2006). Nonetheless, our primary focus

Table 3 Descriptive statistics of numerical features

Feature	Abbr.	Mean	SD	Percentile 25 %	Percentile 75 %	Missing (%)
Diastolic Blood Pressure	DBP	61.0	10.9	53.6	67.2	1.2
Systolic Blood Pressure	SBP	119.5	16.9	107.37	129.70	1.2
Mean Blood Pressure	MBP	78.28	11.13	70.61	84.71	1.21
Glasgow Coma Scale Total	GCST	12.58	2.98	10.60	15.00	2.57
Glucose	GLU	139.35	44.83	111.67	155.25	0.78
Heart Rate	HR	85.59	15.45	74.74	95.52	1.2
Respiratory Rate	RR	18.67	4.01	15.88	20.83	1.3
Oxygen Saturation	OS	96.96	2.39	95.90	98.51	0.77
Temperature	Temp	36.84	0.67	36.46	37.21	2.73
Weight	Weight	81.90	23.72	66.00	94.00	32.19
pH	pH	7.08	0.64	7.06	7.40	30.16

Missing (%) describes the percentage of missing values of the respective feature

lies on the AUROC metric due to its well-established and intuitive nature, enabling a more straightforward interpretation and comparison with existing literature.

To further address the challenge of highly imbalanced targets during model training, we also experimented with data balancing techniques, specifically random under-sampling and synthetic minority over-sampling (SMOTE). However, the application of these methods resulted in a decrease in model performance for threshold-independent metrics such as AUROC. For a detailed presentation of these results, we refer the reader to Appendix C.

We perform hyperparameter optimization for both interpretable and black-box models, with details and the hyperparameter grid provided in Appendix D. Hyperparameter optimization is essential for a fair comparison, as it ensures that each model is tuned to its optimal performance, allowing a more accurate comparison. All models are trained on a workstation equipped with an Nvidia A6000 GPU, Intel i7-12700K (12 cores) CPU, and 128 GB of memory.

3.4 Feature selection for parsimonious models

Constructing parsimonious ML models, especially in high-dimensional data scenarios like our ICU case study, relies on compact feature sets. These sets, composed of the most relevant features, streamline models by reducing computational complexity and enhancing interpretability. Several methods exist for obtaining such compact feature sets, including L1 regularized logistic regression (Lasso regression), forward-selection, and backward-selection (James et al., 2013). L1 regularization, for instance, adds a penalty term to the logistic regression loss function, effectively promoting feature sparsity by driving less critical feature coefficients to zero. Forward-selection gradually incorporates features based on their predictive power, while backward-selection starts with all features and eliminates the least significant ones. These approaches facilitate the selection of a compact set of informative features, thus improving model interpretability.

In this study, we investigate the impact of reducing the feature space on model performance. Therefore, we utilized two types of feature selection: First, we made an informed choice and manually selected the mean of each feature within the 24-hour time period, resulting in a

highly interpretable feature set of 11 features excluding the sensitive features gender, age, and ethnicity. Second, we employed Sequential Forward Floating Selection (SFFS) in conjunction with logistic regression to obtain a compact feature set (Pudil et al., 1994). SFFS is a hybrid feature selection method that combines the strengths of forward- and backward-selection. The SFFS algorithm is outlined in Algorithm 1.

Algorithm 1 Sequential Forward Floating Selection (SFFS)

```

1: Input: Input feature set  $\mathcal{X}$ , criterion function  $J$ , number of features  $k_{max}$ 
2: Output: Selected feature set  $\mathcal{S}$ 
3: Initialize the current set of selected features  $\mathcal{S}$  and the possible set of features  $\mathcal{U}$ 
4:  $\mathcal{S} \leftarrow \emptyset$ 
5:  $\mathcal{U} \leftarrow \mathcal{X}$ 
6: while  $|\mathcal{S}| < k_{max}$  do
7:    $i^* \leftarrow \arg \max_{i \in \mathcal{U}} J(\mathcal{S} \cup i)$                                  $\triangleright$  Select best feature
8:    $\mathcal{S} \leftarrow \mathcal{S} \cup i^*$                                           $\triangleright$  Add  $i^*$  to selected features
9:    $\mathcal{U} \leftarrow \mathcal{U} \setminus i^*$                                           $\triangleright$  Remove  $i^*$  from possible feature set
10:   $P_{\mathcal{S}} \leftarrow J(\mathcal{S})$                                           $\triangleright$  Compute performance on  $\mathcal{S}$ 
11: repeat
12:    $j^* \leftarrow \arg \max_{j \in \mathcal{S}} J(\mathcal{S} \setminus j)$                                  $\triangleright$  Temp. remove feature (floating)
13:    $P_{\mathcal{S} \setminus j^*} \leftarrow J(\mathcal{S} \setminus j^*)$                                           $\triangleright$  Compute performance on  $\mathcal{S} \setminus j^*$ 
14:   if  $P_{\mathcal{S} \setminus j^*} > P_{\mathcal{S}}$  then
15:      $\mathcal{S} \leftarrow \mathcal{S} \setminus j^*$                                           $\triangleright$  Remove feature permanently
16:   end if
17: until  $P_{\mathcal{S} \setminus j^*} \leq P_{\mathcal{S}}$ 
18: end while

```

The algorithm starts with an empty feature set and iteratively incorporates the most significant feature not yet included. Following each addition, the algorithm assesses potential improvements by temporarily removing features from the current set. If this removal leads to a better model, the feature is permanently excluded. This floating process allows SFFS to explore a broader solution space compared to traditional forward- or backward-selection methods, resulting in a more optimal and informative feature subset. As the criterion function J , we employ a 5-fold cross-validation using logistic regression. This selection process is repeated for each classification task individually. It is important to note that we do not use model-specific feature selection. Instead, we derive a uniform set of features for each prediction task using the described method. These feature sets are used across all models, ensuring a fair comparison and allowing us to examine the effect of reducing the feature space on model performance.

Eventually, Table 4 presents an overview of our feature sets. Specifically, we evaluate model performances on our three classification tasks: mortality, LOS3, and LOS7. For each classification task, we consider three options to define feature sets. We either select all features (no selection method), only use 11 selected features, or combine the selected features with the sensitive features gender, age, and ethnicity, for a total of 14 features. The feature selection is done either manually (selecting mean values) or automated using SFFS. Note that the automated feature selection yields three different feature sets for each task, which are detailed in Appendix E.

Table 4 Overview of feature sets

Index	Number of features	Selection method	Sensitive features
$\mathcal{D}_{462-None}$	462	None	
$\mathcal{D}_{465-None-Sens}$	465	None	✓
\mathcal{D}_{11-Man}	11	Manual	
$\mathcal{D}_{11-Auto}$	11	Automated	
$\mathcal{D}_{14-Man-Sens}$	14	Manual	✓
$\mathcal{D}_{14-Auto-Sens}$	14	Automated	✓

Features were selected by the naive manual or the automated algorithmic approach. Sensitive features were either included or excluded. Note, that the reduced feature sets found by SFFS differ between tasks (see Appendix E)

3.5 Mathematical formulation

We consider a set of n training samples comprising input features $x^j \in \mathcal{X}$ and targets $y^j \in \mathcal{Y}$, with $j = 1, \dots, n$. The input space typically is high-dimensional, determined by m features

$$\mathcal{X} = \mathcal{X}^{(1)} \times \dots \times \mathcal{X}^{(m)}. \quad (3)$$

The objective is to train an ML model $\mathcal{M} \in \mathcal{H}$, $\mathcal{H} = \{\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}\}$, by minimizing the empirical risk

$$\min_{\mathcal{M} \in \mathcal{H}} \frac{1}{m} \sum_{j=1}^n \mathcal{L}(y^j, \mathcal{M}(x^j)). \quad (4)$$

Here, $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ represents a loss function that measures the error between target features and training sample predictions. The function space \mathcal{H} , commonly referred to as the hypothesis space, produces an output $\mathcal{M}(x^j)$, also called the prediction.

3.5.1 Generalized additive models

GAMs compute their predictions by summing the outputs of n functions, where n equals the number of input features (Hastie & Tibshirani, 1987; Lou et al., 2012). Thereby, each of these functions $f_i \in \mathcal{F}_{\text{GAM}}$, maps from $\mathcal{X}^{(i)}$ to \mathcal{Y} , where $\mathcal{X}^{(i)}$ denotes the input space of the i -th input feature. The univariate mappings between individual features and the response are called shape functions (Lou et al., 2012). The function space \mathcal{F}_{GAM} is commonly pre-set, ranging from smooth splines to step functions.

GAM-splines. Traditionally, GAM shape functions were learned via splines (Hastie & Tibshirani, 1987). Splines provide a flexible, non-parametric way to model the relationship between individual features and the target. The idea behind using splines in a GAM is to capture non-linear relationships while preventing overfitting from excessive model flexibility. Mathematically, a spline is represented by a linear combination of basis functions, with the coefficients of the linear combination estimated from the data. The most common types of basis functions used in GAMs are cubic splines,

$$f_i(x_i) = \sum_j \beta_{ij} B_j(x_i), \quad (5)$$

where $f_i(x_i)$ is the shape function representing the relationship between the i -th feature x_i and the target. $B_j(x_i)$ are the cubic basis functions evaluated at x_i , and β_{ij} are the coefficients estimated from the data, specific for each feature and each basis function.

Explainable boosting machine (EBM). In more recent approaches bagged and boosted tree ensembles (Lou et al., 2012, 2013) are used. Here, shape functions come from the space of step functions, which are computed as a sequence of shallow trees. Mathematically, the shape functions take the form

$$f_i(x_i) = c_k \mathbb{1}_{x_i \in \mathbf{C}_k} = \begin{cases} c_1 & \text{if } x_i \in \mathbf{C}_1, \\ c_2 & \text{if } x_i \in \mathbf{C}_2, \\ \dots \\ c_K & \text{if } x_i \in \mathbf{C}_K, \end{cases} \quad (6)$$

where $\mathbf{C}_1, \dots, \mathbf{C}_K$ divide the input space $\mathcal{X}^{(i)}$ into K disjoint areas, i.e.,

$$\bigcup_{j=1}^K \mathbf{C}_j = \mathcal{X}^{(i)}, \quad (7)$$

$$\mathbf{C}_j \cap \mathbf{C}_l = \emptyset, \quad j, l = 1, \dots, K, \quad (8)$$

and the scalar c_k denotes the effect features have in their respective areas.

Interpretable generalized additive neural network (IGANN). IGANN, a novel ML model belonging to the GAM family, fits shape functions using a boosted ensemble of neural networks, where each network represents an extreme learning machine (Kraus et al., 2024b). An extreme learning machine is a special type of neural network with a single hidden layer; the input-to-hidden layer weights are randomly assigned and fixed, while only the weights from the hidden layer to the output are optimized during training. This method can improve training time and reduces overfitting (Huang et al., 2006). The shape functions of an IGANN model are expressed as

$$f_i(x_i) = \sum_j W_{ij}^{(2)} \sigma \left(W_{ij}^{(1)} x_i + b_{ij}^{(1)} \right) + b_{ij}^{(2)}, \quad (9)$$

where $W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}$ are weights and biases, and σ is an activation function like Rectified Linear Units (ReLU) (Nair & Hinton, 2010). IGANN uses a sequence of networks, where each element $j \in J$ corresponds to an extreme learning machine.

3.5.2 Black-box and traditional interpretable models

In order to evaluate the performance gap between interpretable GAMs and non-interpretable black-box models, we selected black-box representatives known for their strong predictive capabilities. Additionally, we included traditional interpretable models to provide a more comprehensive comparison.

Random forest (RF). RF is an ensemble learning method that combines multiple decision trees to enhance prediction accuracy (Breiman, 2001). The algorithm constructs each tree independently by recursively splitting the data based on the most informative features. To prevent overfitting, bootstrapping and random feature selection are employed. However, the interpretability of RF can be limited due to the complexity of the ensemble.

eXtreme Gradient Boosting (XGB). XGB is a scalable, efficient, and distributed gradient boosting framework that constructs an ensemble of decision trees to minimize a given loss

function (Chen & Guestrin, 2016). Unlike RF, XGB expands the ensemble by adding new trees that correct the errors made by previous trees. Interpretability of XGB can be challenging due to the complex interactions between features learned by the ensemble.

Decision tree (DT). DT is a category of interpretable ML models that recursively divide the input space into regions based on input feature values (Breiman et al., 1984). Each internal node of the tree represents a feature and its split value, while each leaf node signifies the predicted class (for classification) or value (for regression).

Logistic regression (LR). LR is a straightforward and interpretable linear model used for binary classification. It estimates the probability of an instance belonging to a specific class by applying the logistic function to a linear combination of input features. The model's coefficients can be interpreted as feature importance, offering insights into the relationship between the features and the predicted outcome.

3.6 Model interpretation

In order to demonstrate the advantages of interpretable GAMs, we proceed with a qualitative analysis of the so-called shape plots along three lines. First, we analyse and compare the shape plots of different GAMs, highlighting the differences that arise from the underlying mechanisms (see Sect. 3.5.1). The aim is to identify patterns and relationships that underly the predictive process of the model. In a practical setting, this graphical analysis enables clinicians to understand the behaviour of the model, facilitating the recognition of patterns and the assessment of the plausibility of the learned relationships between features and outcomes. Second, we assess the interpretability and plausibility of our ML models by comparing the shape functions derived from our analysis with existing medical literature. This step is crucial, as counter-intuitive shape functions are concerning to clinicians, and may prevent the potential use of such an ML model. Finally, we consult medical experts to discuss our findings, which serves as a crucial step in verifying the plausibility of our interpretable ML models and ensuring their relevance and applicability in real-world healthcare situations. More details about these consultations can be found in Appendix G.

4 Results

In this section, we showcase the results of our experiments. First, we compare the performance of interpretable and black-box models on the three prediction tasks using all available data. Second, we assess the performance gap between parsimonious ML models using only few features, and their counterparts trained on the full dataset, emphasizing the influence of feature selection methods on model performance. Moreover, we examine the difference in performance resulting from the inclusion or exclusion of sensitive features. Finally, we present selected shape plots to underscore the distinctions between the GAMs and discuss the plausibility of the visualized relationships.

4.1 Model performance on full dataset

Table 5 summarizes the model performances on the complete dataset including sensitive features ($\mathcal{D}_{465-None-Sens}$) and shows the average AUROC and AUPRC values, as well as the standard deviation of these metrics over the folds of the cross-validation. XGB stands out as the model with the highest predictive performance overall, consistently surpassing

other models in both AUROC and AUPRC across all tasks and achieving the highest average ranking.

EBM emerges as the top performer within interpretable models, demonstrating its competitiveness with black-box models. The performance gap between the top black-box model (XGB) and the top GAM (EBM) is minimal, with differences ranging from 0.2 to 0.9 percentage points in AUROC scores and 1.0 to 2.8 percentage points in AUPRC scores across tasks. Notably, EBM not only ranks highest among interpretable models but also outperforms the second black-box model, RF, in all cases except LOS3. This observation highlights the potential of interpretable models, indicating that they can achieve performance comparable to more complex black-box models.

Comparing our results with other benchmark studies on the MIMIC-III dataset, our AUROC scores are consistent with the literature. For instance, Wang et al. (2020) report an AUROC score of 88.7 for LR and 89.7 for RF on the mortality task, while our LR and RF models achieve AUROC scores of 85.3 and 86.3, respectively. Similarly, Harutyunyan et al. (2019) report an AUROC score of 84.8 for the same task, differing by predicting on the first 48 h of patient data, whereas both our study and Wang et al. (2020) predict on the first 24 h. Additionally, Hegselmann et al. (2020) reported results closely resembling ours for predicting mortality using LR, including an EBM model which exhibited comparable performance with an AUROC score of 87.2, aligning with our value of 87.1. For the length-of-stay tasks, our results slightly outperformed those reported by Wang et al. (2020). Additionally, different preprocessing strategies among various studies, such as those by Purushotham et al. (2018), lead to variations in the inclusion or exclusion of specific variables.

In conclusion, the results in Table 5 provide a comprehensive understanding of the strengths and weaknesses of various models using the full dataset. The narrow performance gap between the top black-box and interpretable models encourages further exploration of interpretable models in operations research, particularly in contexts where model interpretability is crucial for decision-making.

4.2 Sequential forward floating selection

Figure 2 displays the results for the sequential forward floating selection approach using logistic regression. The dashed horizontal lines indicate the model performances when utilizing 462 features, excluding the sensitive features gender, age, and ethnicity ($\mathcal{D}_{462-None}$). The results uncover three insights. First, it is remarkable that a single feature can predict mortality, LOS3, and LOS7 with AUROC values ranging from 69.6 to 75.2. Second, only 14 features are needed to achieve a logistic regression model that performs just 1.4 to 4.6 percentage points below the full model trained on 462 features. Third, for all three prediction tasks, a parsimonious linear model trained on a subset of features, excluding sensitive features (gender, age, and ethnicity), can outperform the model trained on 462 features, also excluding these sensitive features, in terms of cross-validation performance. This can be seen by the dotted line, representing the performance of the parsimonious model, crossing the dashed horizontal line, which represents the performance of the model trained on 462 features, excluding sensitive features.

4.3 Model performance on compact feature sets

We now examine the performance of different interpretable and black-box models on compact feature sets obtained via manual selection and SFFS. The results are presented in Tables 6 and

Table 5 Comparison of model performance on all features

	Mortality		LOS > 3 Days		LOS > 7 Days	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
RF	86.3 (0.8)	50.4 (2.0)	78.1 (0.6)	70.6 (0.9)	81.5 (1.3)	42.0 (1.9)
XGB	88.0 (0.5)	53.6 (1.6)	78.8 (0.7)	71.6 (0.8)	82.3 (1.4)	42.1 (1.7)
LR	85.3 (0.8)	45.9 (2.0)	76.4 (0.4)	66.9 (1.0)	80.6 (0.6)	38.2 (0.9)
DT	78.0 (0.7)	34.9 (1.0)	74.0 (0.6)	65.2 (0.7)	78.0 (0.7)	34.9 (0.9)
EBM	87.1 (0.6)	50.8 (1.6)	77.9 (0.7)	69.8 (0.8)	82.1 (0.9)	41.1 (1.5)
GAM-Splines	84.6 (0.8)	48.6 (1.4)	77.1 (0.5)	68.4 (0.6)	81.0 (0.5)	38.8 (0.9)
IGANN	85.8 (0.5)	46.7 (1.1)	76.9 (0.6)	67.8 (0.8)	80.7 (0.5)	38.1 (0.7)
Δ	0.9	2.8	0.9	1.8	0.2	1.0

Sensitive features gender, age, and ethnicity are included. Δ denotes the difference between the best-performing black-box model and the best-performing interpretable model

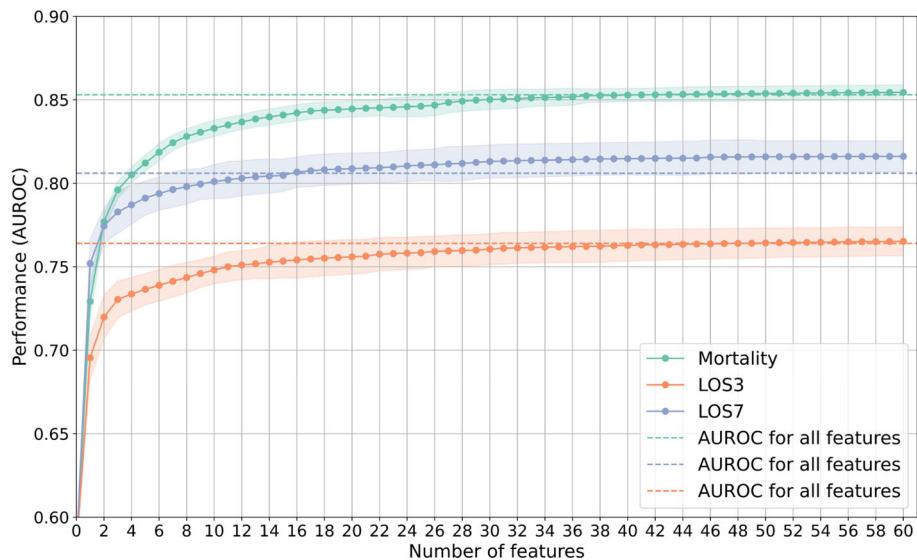


Fig. 2 5-Fold sequential forward floating selection using optimized logistic regression as estimator. The dashed horizontal lines represent the model performances when using 462 features, excluding the sensitive features gender, age, and ethnicity ($\mathcal{D}_{462\text{-}None}$)

7, comparing the performance of models using only 11 (excluding sensitive features) and 14 features (including sensitive features). The numbers in parentheses indicate the performance improvements or deteriorations compared to models trained on the full dataset including sensitive features ($\mathcal{D}_{465\text{-}None\text{-}Sens}$).

Table 6 shows that the performance of all models typically declines as the number of features is reduced to 14 (including sensitive features). However, the performance decrease is not substantial relative to the reduction in the number of features used in the models. For instance, the AUROC for predicting mortality for GAMs and black-box models dropped by 1.4 to 2.7 percentage points despite reducing the feature set by 451 features. This observation suggests that the feature reduction did not significantly impact the overall predictive power of these models. Similar results were obtained for other tasks.

Comparing the performance of models on datasets with different feature selection methods reveals that models generally perform better on datasets with features selected using SFFS than those selected using the mean values of numerical features. While SFFS appears to be a more effective method for feature selection in our setting, it may come at the cost of interpretability, as features based on more complex statistics (e.g., skewness and standard deviation) are introduced (see Table 11 in Appendix E for full feature lists). Consequently, the level of abstraction required to draw conclusions about the implications of a shape plot might be considerably more complex than when using the mean (see Sect. 4.4).

Table 7 presents the performance of the models on reduced datasets with 11 features, excluding the sensitive features. The results indicate that these models achieve comparable performance to those trained with sensitive features included, demonstrating their ability to maintain a satisfactory performance level despite the exclusion. Additionally, it is evident that the performance of the models on the mortality task particularly suffers from the removal of sensitive features.

Table 6 Comparison of model performance on reduced dataset with 14 features, where features were selected using mean value of each feature ($\mathcal{D}_{14-Man-Sens}$) and performing SFFS ($\mathcal{D}_{14-Auto-Sens}$)

	Mortality		LOS > 3 Days		LOS > 7 Days	
	AUROC	AUDRC	AUROC	AUPRC	AUROC	AUDRC
<i>Features selected using the mean values</i>						
RF	84.7 (−1.6)	45.1 (−5.3)	74.3 (−3.8)	65.1 (−5.5)	77.7 (−3.8)	34.8 (−7.2)
XGB	85.8 (−2.2)	46.9 (−6.7)	74.2 (−4.6)	65.0 (−6.6)	78.3 (−4.0)	36.0 (−6.1)
LR	80.2 (−5.1)	37.8 (−8.1)	71.5 (−4.9)	60.9 (−6.0)	75.5 (−5.1)	32.7 (−5.5)
DT	75.9 (−2.1)	31.3 (−3.6)	70.8 (−3.2)	59.8 (−5.4)	75.5 (−2.5)	30.9 (−4.0)
EBM	84.8 (−2.3)	44.6 (−6.2)	73.6 (−4.3)	63.0 (−6.8)	78.1 (−4.0)	35.0 (−6.1)
GAM-Splines	83.3 (−1.3)	42.3 (−6.3)	72.8 (−4.3)	62.2 (−6.2)	76.7 (−4.3)	33.9 (−4.9)
IGANN	83.9 (−1.9)	43.7 (−3.0)	73.0 (−3.9)	62.4 (−5.4)	77.0 (−3.7)	34.2 (−3.9)
Δ	1.0	2.3	0.7	2.1	0.2	1.0
<i>Features selected using sequential forward floating selection</i>						
RF	84.4 (−1.9)	45.0 (−5.4)	76.3 (−1.8)	68.7 (−1.9)	81.0 (−0.5)	39.1 (−2.9)
XGB	85.3 (−2.7)	45.7 (−7.9)	76.8 (−2.0)	69.2 (−2.4)	81.2 (−0.9)	39.1 (−3.0)
LR	83.5 (−2.3)	42.3 (−3.6)	75.0 (−1.4)	64.6 (−2.3)	76.0 (−4.6)	33.2 (−5.0)
DT	78.6 (−0.6)	34.4 (−0.5)	73.1 (−0.9)	64.0 (−1.2)	77.7 (−0.3)	35.9 (+1.0)
EBM	85.0 (−2.1)	45.4 (−5.4)	75.8 (−2.1)	69.8 (−2.8)	80.8 (−1.3)	38.2 (−2.9)
GAM-Splines	84.6 (−)	44.7 (+3.9)	75.4 (−1.7)	66.6 (−1.8)	80.3 (−0.7)	37.7 (−1.1)
IGANN	84.4 (−1.4)	44.5 (−2.2)	75.5 (−1.4)	66.6 (−1.2)	80.4 (−0.3)	38.2 (+0.1)
Δ	0.3	0.3	1.0	0.6	0.4	0.9

Numbers in parentheses denote the improvements (+) or deterioration (−) in model performance compared to the model performance trained on the full dataset. Sensitive features were included. Δ denotes the difference between the best-performing black-box model and the best-performing interpretable model

Table 7 Comparison of model performance on reduced dataset with 11 features (\mathcal{D}_{11-Man})

	Mortality		LOS > 3 Days		LOS > 7 Days	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
<i>Features selected using the mean values</i>						
RF	83.4 (−1.3)	42.1 (−2.0)	73.9 (−0.4)	64.8 (−0.3)	78.0 (−0.3)	35.2 (−0.4)
XGB	84.3 (−1.5)	44.4 (−2.5)	73.9 (−0.3)	64.5 (−0.5)	78.1 (−0.2)	35.5 (−0.5)
LR	79.0 (−1.2)	36.0 (−1.8)	71.3 (−0.2)	60.7 (−0.2)	73.6 (−1.9)	30.9 (−1.8)
DT	76.2 (−0.3)	30.7 (−0.6)	70.8 (−)	59.8 (−)	75.4 (−0.1)	30.9 (−)
EBM	83.7 (−1.1)	42.5 (−2.1)	73.4 (−0.2)	63.0 (−)	78.0 (−0.1)	34.9 (−0.1)
GAM-Splines	82.2 (−1.1)	41.3 (−1.0)	72.8 (−)	62.2 (−)	73.2 (−3.5)	31.4 (−2.5)
IGANN	82.7 (−1.2)	41.8 (−1.9)	72.9 (−0.1)	62.5 (−0.1)	77.1 (−0.1)	34.3 (−0.1)
Δ	0.6	1.9	0.5	1.5	0.1	0.6
<i>Features selected using sequential forward floating selection</i>						
RF	83.9 (−0.5)	44.1 (−0.9)	76.0 (−0.3)	68.5 (−0.2)	80.7 (−0.3)	38.6 (−0.4)
XGB	84.4 (−0.9)	44.4 (−0.9)	76.5 (−0.3)	68.9 (−0.3)	80.9 (−0.3)	38.8 (−0.6)
LR	82.6 (−0.9)	41.1 (−1.2)	74.8 (−0.2)	64.6 (−)	76.0 (−)	33.2 (−)
DT	78.1 (−0.5)	33.2 (−1.2)	73.0 (−0.1)	63.9 (−0.1)	77.6 (−0.1)	35.3 (−)
EBM	83.9 (−1.1)	43.8 (−1.6)	75.7 (−0.1)	66.9 (−2.9)	80.7 (−0.1)	38.1 (−0.1)
GAM-Splines	83.6 (−1.0)	43.7 (−1.0)	75.2 (−0.2)	66.3 (−0.3)	80.2 (−0.1)	37.5 (−0.2)
IGANN	83.5 (−0.9)	43.2 (−1.3)	75.2 (−0.3)	66.4 (−0.2)	80.2 (−0.2)	37.9 (−0.3)
Δ	0.5	0.6	0.8	2.0	0.2	0.7

Sensitive features were excluded. Numbers in parentheses denote the improvements (+) or deterioration (−) in model performance compared to the model including sensitive features. Δ denotes the difference between the best-performing black-box model and the best-performing interpretable model

Tables 6 and 7 compare the performance of black-box and interpretable models across three tasks (mortality, LOS3, LOS7). The best performing black-box model, XGB, shows slightly higher AUROC scores than the best performing interpretable model, EBM, for all tasks and both 11- and 14-feature datasets, with EBM's AUROC scores only 0.1 to 1.0 percentage points lower. This suggests that EBM offers competitive performance while providing interpretability. Compared to the second black-box model, RF, EBM even shows favorable results in three out of twelve settings and shows equal results in another three.

Overall, the black-box and interpretable models showed comparable performance on compact feature sets. The results indicate that GAMs are capable of performing well on compact feature sets obtained using the two feature selection techniques. Furthermore, the performance losses are task-dependent when the sensitive features are excluded from the feature set. A key observation is that performance differences between full and reduced datasets typically exceed differences between black-box and interpretable models. This is also true for differences between manual and automated feature selection. The pattern continues for comparisons between datasets with and without sensitive features in the mortality tasks. This suggests that feature selection has a greater impact than the choice between black-box and interpretable models.

4.4 Assessment of shape plots

In this section, we present the visual output of the interpretable models. By analyzing examples of the generated shape functions, we discuss their plausibility and interpretability. To evaluate the plausibility of the shape plots, we compared the displayed effects of the features, when feasible, with the effects suggested by widely used scoring systems such as SAPS or APACHE.

Additionally, we consulted with multiple medical experts (MEs) and discussed the plots presented in this section to ensure their correctness and alignment with established medical expertise. For more information about the meetings and the medical experts' background, we refer the interested reader to Appendix G as well as to Appendix H, which contains sample quotes from the interviews with the medical experts that illustrate how we arrived at our findings.

It should be noted that the choice of data extraction and preprocessing techniques – such as outlier handling, imputation, and data balancing – along with the selected feature set can influence the resulting shape plots. We focus our assessment on these plots generated with the parameters outlined in Sect. 3.3 and using models trained on the reduced datasets. However in the following figures, shaped curves illustrate the impact of numerical features, while the impact of categorical features is depicted by bar charts. In all plots the x-axis represents the feature value, and the y-axis indicates the influence on the target (in log odds), with positive influence for $y > 0$ and negative influence for $y < 0$.

4.4.1 Assessment of shape plots from manually selected features

Figure 3 examines the impact of patients' age on three prediction tasks and various models, illustrating shape plots for the different GAMs and logistic regression. Columns represent models and rows represent prediction tasks. All three GAMs display similar trends, suggesting data correlations. However, distinctions emerge. For instance, EBM generates noisy functions with abrupt transitions, GAM-Splines produce less frequent, large fluctuations, and IGANN yields smooth curves. The sharp jumps in EBM's shape plot may seem confusing

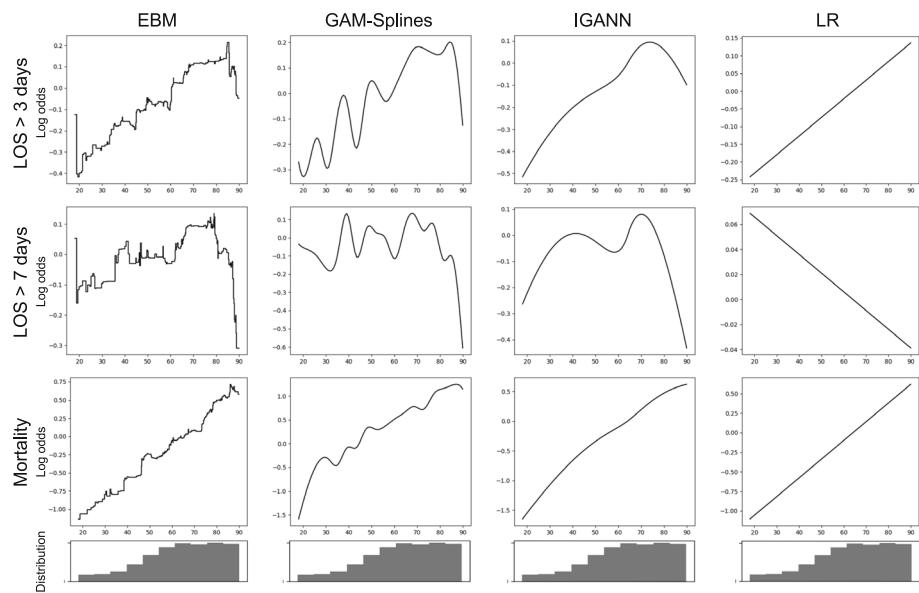


Fig. 3 Shown is the effect of the age feature on different prediction targets. Within each plot, age (in years) is on the x-axis and the effect on prediction (in log odds) is on the y-axis (higher values indicate a higher probability of mortality, LOS3 or LOS7). The grid contains three rows of shape plots, one for each prediction task. The 4 columns represent different GAMs and LR. The bottom row shows how the age feature is distributed in the sample. All plots were created after training the models on the manually reduced dataset ($\mathcal{D}_{14\text{-Man-Sens}}$)

(ME1, ME2) for continuous features like age but are suitable for identifying thresholds in feature spaces. Also, features with very narrow ranges, such as temperature, can be better analysed using more abrupt transitions, such as EBM (ME3). Conversely, the extremely smooth curves produced by IGANN are easier to read (ME1), but thresholds might remain hidden or small, and meaningful fluctuations may go unnoticed.

All models exhibit a near-monotonic relationship between age and mortality, aligning with assessment tools like SAPS (Moreno et al., 2005). The relationship between age and LOS3 shows a similar trend. However, a sharp decrease in probability occurs for ages above 80, suggesting older patients have shorter ICU stays, potentially due to early death or transfer to palliative care units. This assumption is supported by the more pronounced effect for LOS7. The linear model fails to capture this trend reversal between the input feature and the target, indicating that a linear assumption may not adequately represent the relationship between age and length-of-stay. This example highlights situations where simpler, linear models may be less appropriate for describing feature effects and hinder interpretability.

Figure 4 presents a set of shape plots for predicting mortality on the manually reduced dataset with sensitive features. All three GAMs consistently link abnormal temperature and blood pressure to higher mortality rates. Both upward and downward deviations are harmful, with the latter being particularly detrimental. This is consistent with SAPS scores (Moreno et al., 2005) and the experience of the medical experts. An unexpected relationship arises between mortality and Glasgow Coma Scale Total (GCST) score: a monotonic negative relationship is expected, as patients with higher GCST scores are classified as more conscious. Although this negative trend is recognizable, all GAMs show a steep increase in mortality probability between 13 and 14, contradicting medical experts (ME1) as well as medical

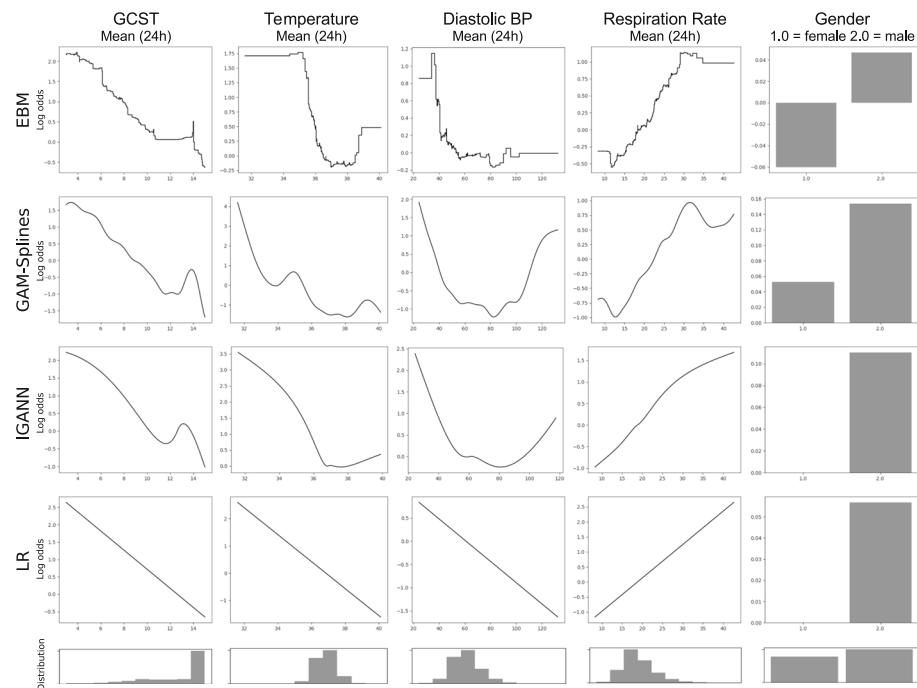


Fig. 4 Shown is the effect of 5 (out of 14) features on patient mortality. We used the manually generated reduced dataset including sensitive features. In each plot, different values of the features are on the x-axis and the effect on mortality (in log odds) is on the y-axis (higher values indicate a higher probability of mortality). The grid contains 4 rows, one for each model, and 5 columns, each representing a feature. The bottom row shows the feature distribution in the sample. All plots were created after training the models on the reduced dataset ($\mathcal{D}_{14\text{-Man-Sens}}$)

literature (Teasdale & Jennett, 1974). Furthermore, all models indicate higher mortality for males, a discrepancy debated but unconfirmed in medical literature (Hollinger et al., 2019). Appendix F shows the shape plots for the remaining features.

4.4.2 Assessment of shape plots from automatically selected features

Finally, we consider a selection of shape plots for predicting mortality based on the reduced dataset generated with SFFS, excluding sensitive features ($\mathcal{D}_{11\text{-Auto}}$). The selected plots are shown in Fig. 5. We observe nearly identical shape plots for diastolic blood pressure and temperature, selected by both SFFS and the manual approach. This also holds for respiratory rate, with SFFS considering only the initial 12 h of ICU stay. For the GCST score, SFFS considers the final 10% of the first 24 h (i.e., the last 144 min), revealing the expected negative monotonic correlation and increasing plausibility. Additionally, a GCST-based feature, the standard deviation of the first 12 h, exhibits a clear negative trend. However, interpreting features based on standard deviation can be challenging, as also ME4 confirms. This negative trend in the feature effect became more understandable after discussions with medical experts who confirmed that it is very common for patients to be admitted to the ICU under anesthesia after a major medical procedure and then wake up as planned (ME1, ME3). A high GCST standard deviation indicates that patients experience a change between comatose and fully

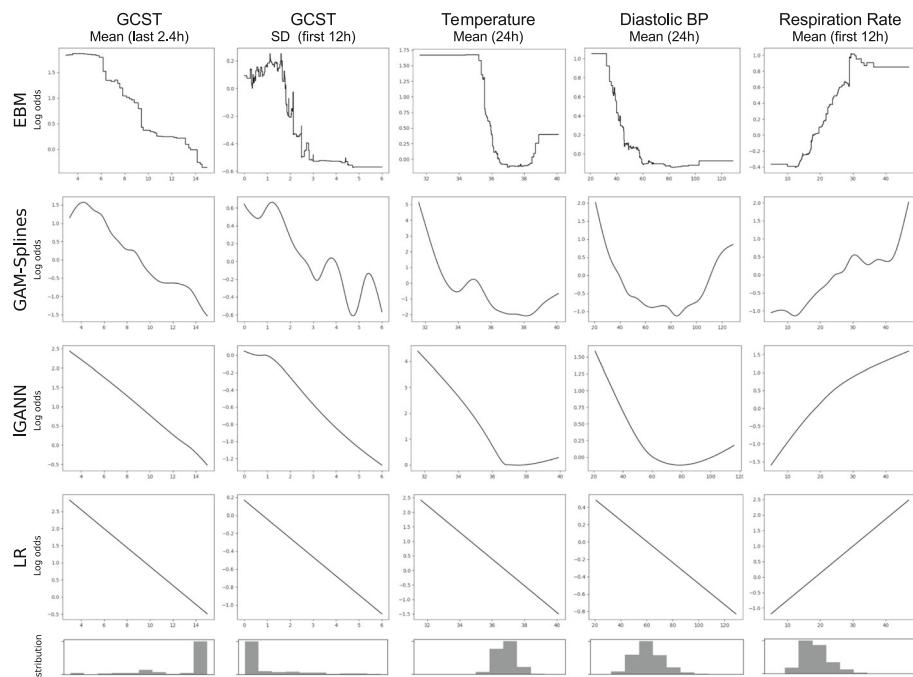


Fig. 5 Shown is the effect of 5 (out of 11) features on patient mortality. This is the reduced dataset generated by SFFS without the sensitive features ($\mathcal{D}_{11\text{-Auto}}$). In each plot, different values of the features are on the x-axis and the effect on mortality (in log odds) is on the y-axis (higher values indicate a higher probability of mortality). The grid contains 4 rows, one for each model, and 5 columns, each representing a feature. The bottom row shows how the features are distributed in the sample

conscious during the first 12h of their ICU stay. However, the function does not indicate the direction of the change. Medical experts express concern about including such logic in a model (ME2, ME3).

5 Discussion

5.1 Interpretation of results

The trade-off between model performance and interpretability is crucial in healthcare analytics and widely discussed in the operations research literature (Yang, 2022; Bertsimas et al., 2021). Stakeholders require a balance between interpretability and performance, focusing on actionable insights (Coussement & Benoit, 2021), regulatory compliance (e.g., GDPR) (Parliament and Council of the European Union, 2016), and ethical fairness (Goodman & Flaxman, 2017).

Our experiments yield several important insights. First, GAMs emerge as a promising solution for addressing the performance-interpretability trade-off in healthcare analytics. The minimal performance gap between GAMs and leading black-box models makes them an attractive option for balancing predictive power and interpretability. This finding aligns with the work of Chang et al. (2021) and Zschech et al. (2022), who demonstrated that performance

penalties associated with interpretable models can vanish entirely in certain settings or even turn into performance advantages.

Our results are consistent with other benchmark studies on the MIMIC-III dataset, as discussed in Sect. 4.1. The AUROC scores we obtained for the mortality task with LR and RF models are comparable to those reported by Wang et al. (2020) and Harutyunyan et al. (2019). Furthermore, our EBM model's performance closely matches the results of Hegselmann et al. (2020) for the same task. These comparisons validate the robustness of our findings and demonstrate that our models' performance aligns with the state-of-the-art in the field. However, it is important to note that directly comparing results across studies can be challenging due to differences in preprocessing strategies and variable inclusion. For instance, Purushotham et al. (2018) employed different preprocessing techniques, leading to variations in the variables used for modeling. These differences highlight the need for caution when making direct comparisons and emphasize the importance of considering the specific context and methodology of each study.

Second, reducing features to obtain parsimonious models or excluding sensitive features can lead to accuracy loss, which in our case was more pronounced with manual approaches than automated methods like Sequential Floating Forward Selection. Notably, differences in predictive performance between feature selection methods were consistently more substantial than disparities between black-box and interpretable models. For instance, sensitive features like gender, age, and ethnicity showed varying importance across prediction tasks. Particularly, for mortality prediction, the impact of excluding these features was much larger than for length-of-stay prediction quality.

Third, shape plot comparisons across dataset versions indicate that GAMs exhibit stability against feature selection procedures. The SFFS-based approach, while resulting in higher agreement with medical evidence, includes more complex features, complicating interpretation. As demonstrated in our study, interpreting features based on standard deviation is challenging, and this complexity increases for features based on skewness or the number of measurements taken. Nonetheless, such features can be strong predictors; thus, when utilizing feature selection methods focused on interpretability, it is essential to balance both predictive capacity and simplicity of feature selection.

Overall, GAMs align well with medical knowledge, but some plot details require further investigation. The learned relationships remain stable despite dataset changes due to sensitive feature exclusion or changes in time periods. While GAMs are desirable for their accuracy and interpretability, it is crucial to note that they are not causal models: shape plots demonstrate prediction generation but do not provide reasons for learned patterns or intervention effects. The implications of these findings will be discussed further in the following sections.

5.2 Practical implications

Our study has several practical implications for the application of GAMs in medical data science and healthcare settings.

Interpretability and collaboration. We reaffirm that GAMs are well-suited for medical data science applications due to their high predictive performance and interpretability (Caruana et al., 2015; Agarwal & Das, 2020). To achieve interpretability, models require careful, manageable, and often time-consuming feature extraction and preprocessing steps. Insufficient preprocessing can lead to ambiguous shape plots, particularly in regions with only few data points or when extreme and implausible values, such as negative age, occur. As a result,

a rigorous process calls for close collaboration between medical experts and data scientists, who must comprehend each other's domains.

Our consultations with medical experts show, that shape plots enable communication between model developers and domain experts. This type of visualizations help to exchange insights and refine understanding of the data and model predictions. Such improved communication can also inform decision-makers about the reasoning behind a model's predictions. The utility of shape plots resides in its capacity to bridge the gap between domain knowledge and model development. Domain knowledge can enhance the accuracy and relevance of models, while the models can offer new insights into the domain, uncovering previously unknown relationships or patterns. By fostering a continuous feedback loop between domain experts and data scientists, shape plots can help ensure that models better align with the needs of the domain, ultimately improving interpretability and real-world applicability.

Our study demonstrates that GAMs offer a balance between performance and interpretability. However, it is worth noting that the interpretability of GAMs may be perceived differently by medical experts, as evident from the expert opinions provided in Appendix H. For instance, ME1 expresses a preference for the smoother, less abrupt transitions in the spline-based GAM, indicating that they would not mind a flatter, slightly averaged graph. In contrast, ME3 suggests that for features with a narrow range, such as temperature, a more precise representation like EBM might be more practical.

These diverse opinions suggest that the interpretability of GAMs may be perceived differently by medical experts, highlighting the potential importance of comprehensibility in healthcare decision-making. In this context, comprehensibility could be understood as the ease with which domain experts can grasp and reason about a model's predictions, which may play a role in the adoption of ML models in clinical settings (Sivaraman et al., 2023). As ME2 points out, the overall trend might be more important than minor fluctuations in the graph, indicating that the level of detail required for comprehensibility may vary depending on the specific use case and the preferences of the medical experts involved.

Consequently, when considering interpretable models for healthcare applications, it might be beneficial to take into account not only the predictive performance but also the level of comprehensibility offered by the model, as perceived by the intended users. This could be explored through close collaboration with domain experts, as demonstrated in our study. By engaging in a dialogue with medical professionals and considering their diverse perspectives, we may be better positioned to develop models that are not only accurate but also more easily understood and accepted by healthcare practitioners, potentially leading to better integration of ML models in clinical decision support systems. However, further research is needed to establish the extent to which comprehensibility influences the adoption and effective use of ML models in healthcare settings.

Ethical considerations. It is important to note that our study focuses on improving predictive performance using interpretable and parsimonious ML models, while ethical considerations are neglected. However, during our consultations various aspects have been discussed which showcase the complexity of deploying ML models in the context of mortality prediction. One critical issue raised involves the ethical implications of deploying predictive models in intensive care settings. As noted by one of our medical experts, the use of algorithms to predict low survival probabilities may not serve the therapeutic goals, especially when it informs clinical decisions about end-of-life care (ME3). The ethical concerns extend to situations where the predictive models might suggest aggressive treatments for patients with a very low chance of survival. This could lead to unnecessary prolongation of suffering, a concern that another expert highlighted while discussing the application of such models in real-world scenarios (ME4).

In a more general medical use case, there is growing concern over the potential unfair and unethical impact of sensitive features, such as gender, age, and ethnicity when applying ML models. As a result, these features are often excluded to avoid these effects. However, simply omitting sensitive features can cause correlated features to partially absorb the effect of the omitted features. This can result in the sensitive feature not being present, but its effect still being at work, hidden in proxy features. While there are no visible effects in the shape plots when such features are removed from our dataset, similar effects have been described in other healthcare settings (Obermeyer et al., 2019).

In general, sensitive features might be reasonable inputs for decision-making in healthcare, but they could also be the source of significant injustices (Vyas et al., 2020). Fully interpretable models, such as GAMs, allow for a close look at the impact of these features, helping prevent unfair decisions. This approach is not possible with black-box models, which remain incomprehensible in their decision logic (Rudin, 2019). Thus, prioritizing interpretability in medical decision-making is crucial to avoid unethical and unfair effects.

ML adoption in healthcare. GAMs can help facilitate ML adoption in healthcare by providing comprehensive, case-by-case explanations for predictions before deployment, addressing ethical and legal concerns. ML models should be designed to complement the expertise of healthcare professionals, rather than replace them entirely (Rajpurkar et al., 2022). In doing so, ML models can provide valuable insights and support for decision-making processes that ultimately lead to better health outcomes for patients.

5.3 Limitations and further research

Despite its contributions, this study presents some limitations and suggests opportunities for further research. One key limitation is the absence of feature interactions in our analysis, which was also confirmed in our consultations, where it was pointed out that physiological systems always interact and therefore modelling interactions is important (ME4). Even though certain GAMs can accommodate second-order interactions and produce heat-map-like shape plots (Lou et al., 2013), incorporating interactions would exponentially increase the models' complexity, as there are $\binom{n}{2} = \frac{n!}{2!(n-2)!}$ potential interaction terms for n features (e.g., for our $n = 465$ features, this would result in 107,880 interaction terms). In addition, incorporating interactions would alter the models so that they no longer conform to the inherent structure of GAMs. In addition, the methods by which models identify interactions vary greatly between different GAM implementations, making it difficult to compare results. To address this limitation, future research could develop more efficient algorithms to detect interactions in a model-agnostic manner (Lengerich et al., 2020). As demonstrated by Topuz et al. (2018) in Bayesian belief networks, one possible approach might involve using domain knowledge, with a physician identifying the most relevant interactions to include. When discussing interactions, the experts emphasize their relevance.

In our experiments with different imputation methods, mean imputation proves to be the best performing. In our extracted dataset, most variables have low missing rates, which limits the impact of imputation. However, it is important to recognize the limitations of mean imputation and its impact on model reliability. These include the potential loss of data variability and the introduction of bias by assuming the same value across different patient populations (e.g., same weight for male and female patients). Further research on datasets with higher missing rates could hold new insights. Similar considerations apply to data balancing techniques. Although our initial experiments with random under-sampling and

SMOTE over-sampling were less promising, future studies may identify the circumstances under which these techniques may be beneficial.

On a technical level, this study utilizes a single method for selecting the optimized feature set based on LR and SFFS. Although this method provides high interpretability and comparability with manual procedures, it might not be ideal, especially for features displaying strong correlation. Investigating a model-agnostic approach focused on relevance and redundancy (Yu & Liu, 2004) could offer a promising alternative, though it may present interpretability challenges.

Lastly, the study's reliance on a single dataset constrains the generalizability of its findings. To enhance generalizability, future research should examine data from various ICUs across different regions. Such analyses could yield valuable insights by comparing shape plots among ICUs with diverse treatment procedures or capabilities.

6 Conclusion

In conclusion, this research contributes to healthcare operations research by examining the use of interpretable models, specifically focusing on GAMs in the ICU context. Our study establishes that GAMs can deliver high performance while preserving interpretability, thus addressing the critical need for transparency in healthcare decision-making. Our findings indicate that feature selection and extraction have a more substantial impact on predictive accuracy than the choice between interpretable and non-interpretable models. Consequently, future research should prioritize identifying best practices for feature selection and development. Additionally, ethical concerns, such as the potential unfair and unethical influence of certain features, must be considered when creating ML models for healthcare. We also underscore the importance of collaboration between medical experts and data scientists in developing accurate and comprehensible models.

Appendix A: Chosen upper and lower bounds to avoid implausible values

As described in Sect. 3.3, we remove implausible values from the time series before calculating summary statistics such as mean, standard deviation, minimum, and maximum. This procedure is based on the approach by Hegselmann et al. (2020) and helps to mitigate errors from the data collection process that could significantly affect the ML models. The upper and lower bounds chosen are listed in Table 8. It is important to note that we do not remove a patient from our dataset when an implausible value occurs; instead, we remove the individual measurement from the time series associated with that patient and then calculate the summary statistics on the remaining data points. Therefore, the lower and upper bounds are chosen to exclude only values that are so far from a regular value that we can be reasonably confident that it must be an error.

Table 8 Established lower and upper bounds for time series preprocessing

Measurement	Lower bound	Upper bound
Glucose	5	2000
Systolic blood pressure	0	400
Temperature	20	45
Weight	20	500
Diastolic blood pressure	20	350
Glasgow coma scale total	3	15
Mean blood pressure	20	400
Heart rate	10	300
Oxygen saturation	20	200
pH	5	9
Respiratory rate	5	50

Appendix B: Experiments with advanced imputation techniques

As described in Sect. 3.3, we explored several imputation techniques beyond simple mean imputation. This section provides a brief overview of our experimental methods and results. The techniques examined include median imputation, k-nearest neighbor (knn) imputation, and two iterative methods: iterative random forest (iRF) and iterative linear regression (iLR). To effectively manage computational requirements, these tests were performed on the dataset containing the 14 manually selected features and the same hyperparameter grid as the other experiments in this study. For all three prediction targets – mortality, LOS3, and LOS7 – our results were consistent. Notably, the more complex imputation methods did not improve performance. For most models with the exception of LR, these techniques actually decreased predictive accuracy. We assume that the small performance losses of complex imputation methods are due to the low incidence of missing values and the indeterminate nature of the missing data mechanism in our dataset. As a result, simple mean imputation emerged as the most robust method, consistently producing better results. This suggests that in datasets characterized by low missing rates and ambiguous missing data mechanisms, simpler imputation techniques may outperform more sophisticated alternatives. In Table 9, we compare the performance for the different imputation techniques and report the performance difference to the mean imputation.

Appendix C: Experiments with data balancing techniques

We explore random under-sampling and Synthetic Minority Oversampling Technique (SMOTE) to address imbalanced target distributions (see section 3.3). The tests are performed on the dataset containing the 14 manually selected features and the same hyperparameters as other experiments. For all prediction targets (mortality, LOS3, LOS7), both techniques decreased predictive accuracy, with more pronounced decreases as more majority samples are removed (under-sampling) or more synthesized minority samples are created (over-sampling). Figure 6 plots the performance for different balancing techniques and ratios, defined as: ratio = $\alpha = \frac{N_{\text{minority}}}{N_{\text{majority}}}$.

Table 9 Comparison of AUROC differences (Δ) for various imputation methods relative to mean imputation for different prediction targets

Model	Mortality		LOS > 3 Days		LOS > 7 Days	
	Method	Δ AUROC	Method	Δ AUROC	Method	Δ AUROC
XGB	Mean	–	Mean	–	Mean	–
	Median	–0.508726	Median	–0.488815	Median	–0.625327
	iLR	–1.403181	iRF	–0.856955	iLR	–0.869722
	iRF	–1.511133	iLR	–0.894452	knn	–0.945321
	knn	–1.567955	knn	–1.149458	iRF	–0.950152
RF	Mean	–	Mean	–	Mean	–
	Median	–0.235323	Median	–0.336272	Median	–0.230433
	iRF	–1.016325	iLR	–0.451807	iLR	–0.505472
	iLR	–1.210919	knn	–0.497799	knn	–0.538565
	knn	–1.254153	iRF	–0.626240	iRF	–0.665823
GAM-splines	Median	0.079474	iRF	0.048688	iRF	0.184649
	Mean	–	iLR	0.007246	knn	0.153558
	iRF	–0.336092	Mean	–	iLR	0.130880
	iLR	–0.442228	knn	–0.126122	Mean	–
	knn	–0.642070	Median	–0.267787	Median	–0.092878
LR	iRF	0.690156	iRF	0.238541	knn	0.375990
	iLR	0.253291	iLR	0.156303	iRF	0.306776
	knn	0.250128	knn	0.133322	iLR	0.265012
	Mean	–	Mean	–	Mean	–
	Median	–0.678840	Median	–0.494478	Median	–0.672530
IGANN	Mean	–	iLR	0.004094	iRF	0.061074
	Median	–0.322005	Mean	–	Mean	–
	iRF	–0.354734	iRF	–0.091654	iLR	–0.042336
	iLR	–0.473279	knn	–0.092008	knn	–0.128729
	knn	–0.619628	Median	–0.442074	median	–0.465625
EBM	Mean	–	mean	–	Mean	–
	Median	–0.370593	Median	–0.448254	median	–0.609137
	iRF	–1.165603	iLR	–0.780038	knn	–0.915631
	iLR	–1.205107	knn	–0.862468	iLR	–1.116673
	knn	–1.391871	iRF	–0.875292	iRF	–1.193140
DT	Mean	–	Mean	–	Mean	–
	Median	–1.115663	Median	–0.430361	knn	–0.964964
	iLR	–1.519647	iRF	–0.568668	iRF	–1.177052
	iRF	–1.564441	knn	–0.625558	iLR	–1.306374
	knn	–1.719530	iLR	–0.694613	median	–1.399801

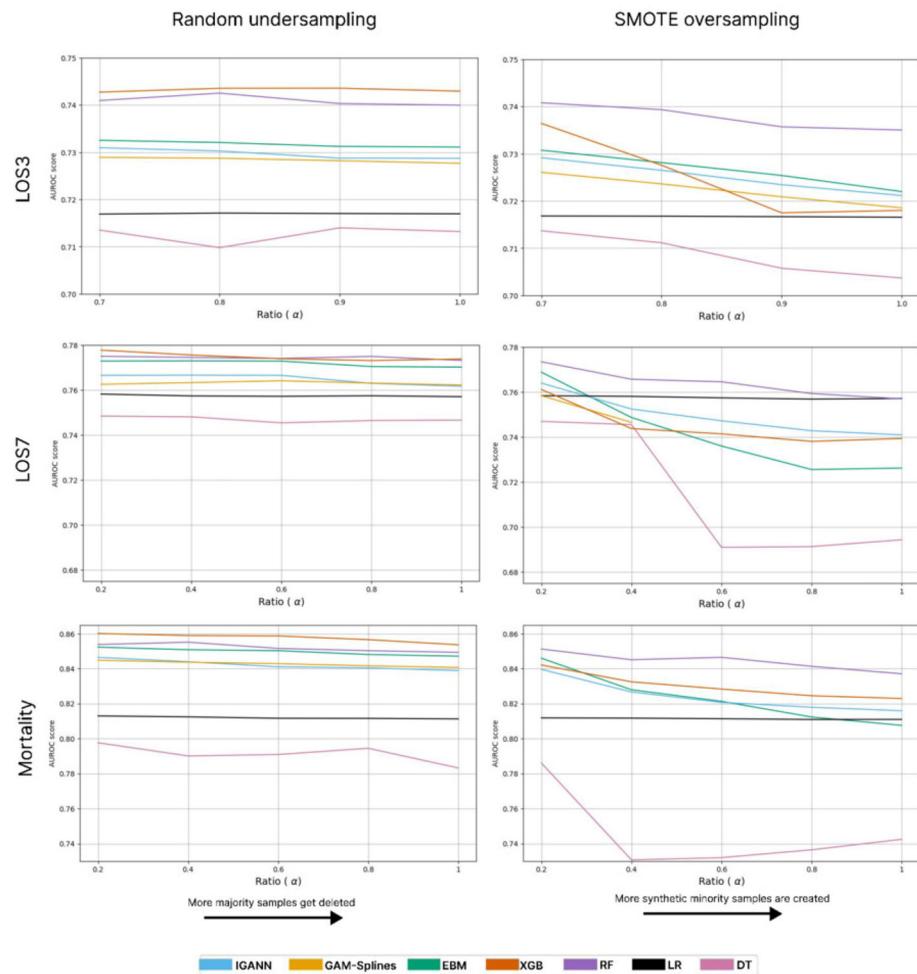


Fig. 6 Results for the three prediction targets with applied data balancing techniques. Left: random undersampling, removing majority samples until a specific ratio α is reached. Right: SMOTE over-sampling, creating synthetic minority samples until a specific ratio is reached

Appendix D: Implementation and hyperparameter tuning

We implemented and evaluated all models in Python. The decision tree, random forest as well as logistic regression are implemented using scikit-learn⁵ version 1.1.2. XGB is implemented using the open-source implementation from xgboost⁶ version 1.6.2. GAM-Splines are implemented using pyGAM⁷ version 0.8.0. EBM is implemented using the the inter-

⁵ <https://scikit-learn.org/>.

⁶ <https://github.com/dmlc/xgboost>.

⁷ <https://github.com/dswah/pyGAM>.

Table 10 Grid showing all tested parameter combinations. Note that the best-performing hyperparameters vary across our tasks

Model	Tuning parameters	Tuning range
XGB	Num. estimators	50, 100, 200, 500, 1000, 2000
	Max. depth	None, 3, 6, 9, 12
	Learning rate	0.01, 0.1, 0.3
Random Forest	Num. estimators	50, 100, 200, 500, 1000
	Max. depth	None, 5, 10, 20, 40
	Class weight	None, balanced
Logistic Regression	Regularization strength	1e-3, 1e-2, ..., 1e2, 1e3
	Penalty term	L1, L2
	Solver	lbfgs, saga
	Class weight	None, balanced
Decision Tree	Max. depth	None, 5, 10, 20, 40
	Max. leaf nodes	None, 5, 10, 20, 40
	Class weight	None, balanced
EBM	Splitter	best, random
	Max. bins	256, 512
	Outer bags	8, 16
GAM-Splines	Inner bags	0, 4
	Num. splines	5, 10, 15, 20, 25
	Regularization strength	Log scale: 10^{-3} to 10^4
IGANN	ELM scaling parameter	1, 2, 5
	Boosting rate	0.025, 0.1

pretML⁸ package version 0.2.7. IGANN⁹ can be obtained from GitHub. We applied a grid search method to systematically explore the hyperparameter space for each model. By testing different combinations of hyperparameters, we aimed to identify the optimal configuration for each model to ensure a fair comparison. Table 10 lists our hyperparameter grid.

Appendix E: Comparison of features selected by the automated vs. the manual approach

We employ two types of feature selection, as discussed in Sect. 3.4. The manually selected feature set (\mathcal{D}_{11-Man}) remains the same across all tasks, while Sequential Forward Feature Selection (SFFS) finds potentially optimal feature sets ($\mathcal{D}_{11-Auto}$ and $\mathcal{D}_{14-Auto-Sens}$) for each task. Table 11 lists the manually and SFFS-selected features for reproducibility and transparency. The feature names are derived as follows: an abbreviation for the time series from which the measurements were taken (e.g., MBP = Mean blood pressure), followed by the specific portion of the time series considered, given as a percentage, including the sign indicating whether the (–) last or (+) first part of the time series is used (e.g., +50% = first 12 h,

⁸ <https://github.com/interpretml/interpret>.

⁹ <https://github.com/MathiasKraus/igann>.

Table 11 Task-specific features selected by either the manual (mean values) or automated (SFFS) approach

Task	Selection method	
	\mathcal{D}_{11-Man}	$\mathcal{D}_{11-Auto}$
Mortality	GCST+100%mean	MBP+100%min
	Weight+100%mean	OS-10%mean
	HR+100%mean	RR+50%mean
	MBP+100%mean	pH+100%std
	SBP+100%mean	GCST-10%mean
	GLU+100%mean	Temp+100%mean
	RR+100%mean	DBP-50%mean
	DBP+100%mean	GCST+50%std
	Temp+100%mean	HR-25%max
	OS+100%mean	GLU-50%min
	pH+100%mean	Weight-10%min
	GCST+100%mean	Ph+50%std
	Weight+100%mean	GCST+100%std
	HR+100%mean	DBP-10%min
LOS > 3 days	MBP+100%mean	GCST-25%len
	SBP+100%mean	RR+100%mean
	GLU+100%mean	GCST-50%mean
	RR+100%mean	OS+100%min
	DBP+100%mean	GCST-10%mean
	Temp+100%mean	pH-25%len
	OS+100%mean	SBP+100%min
	pH+100%mean	HR-10%mean
	GCST+100%mean	GCST-25%len
	Weight+100%mean	MBP+100%min
	HR+100%mean	OS-50%skew
	MBP+100%mean	GCST+100%std
	SBP+100%mean	HR-25%max
	GLU+100%mean	pH-25%len
LOS > 7 days	RR+100%mean	GCST-25%mean
	DBP+100%mean	pH+50%std
	Temp+100%mean	OS+50%mean
	OS+100%mean	RR+100%mean
	pH+100%mean	RR+100%min
	GCST+100%mean	
	Weight+100%mean	
	HR+100%mean	
	MBP+100%mean	
	SBP+100%mean	
	GLU+100%mean	
	RR+100%mean	
	DBP+100%mean	
	Temp+100%mean	

-10% = last 2.4h), and an abbreviation indicating which summary statistic (mean, standard deviation, minimum, maximum, skewness, and number of measurements) is used to calculate the feature (e.g., len = number of measurements).

Appendix F: Additional shape plots

This section presents 36 additional shape plots that complement the plots shown in Fig. 4, displaying the effect of the remaining 9 (out of 14) features on patient mortality risk learned by the GAMs analyzed in Sect. 4.4. These plots provide a more holistic view of the predictive behavior of the models, allowing the reader to review a complete basis for the predictions and gain a better understanding of the overall model behavior.

Appendix G: Consultations with medical experts

To assess the interpretability and alignment of the shape plots with medical knowledge, we consult four medical experts and discuss eight sets of questions (Q0–Q7).

During these consultations, we first ask the experts about their medical backgrounds and experience with ICU patients to establish the context (Q0). We then provide an overview of our study, explaining the rationale for comparing interpretable and black-box ML models in ICU decision-making. Next, we delve into the data basis of our study, detailing our use of the MIMIC-III database. We describe our prediction targets, mortality and length-of-stay, and outline the process of creating a supervised ML model for this purpose. We seek opinions on the relevance of ML predictions, their potential integration into ICU workflows, and associated risks and challenges (Q1). Emphasizing the importance of evaluating ML-generated medical prognoses, we gather expert assessments of the ML system. We then focus on perceptions of ML-based predictions, specifically trust in opaque ML systems and measures to increase trust (Q2). Next, we describe interpretable ML models, particularly GAMs, highlight their transparency and visualization methods. We discuss the process of variable extraction for learning and predictions, addressing the comprehensibility of shape plots with questions about their usefulness and clarity (Q3). We analyze specific features to determine whether the relationships shown are consistent with medical knowledge (Q4). We assess the suitability of such a prediction model for in-hospital use, considering the comprehensibility of shape plots, the number of plots required, and the importance of feature interactions (Q5). We compare different shape plots from various GAMs, looking for preferences and observed differences (Q6). Finally, we discuss an example of a more complex feature and its shape plot interpretation, focusing on their implications for prediction (Q7).

All consultations are conducted face-to-face or via video conference and are supported by a presentation. These sessions are recorded, transcribed, and analyzed for insights regarding the interpretability and alignment of the shape plots with the experts' medical knowledge. Table 12 provides basic information about the experts, including their current positions and experience with intensive care patients. The interested reader can access the translated version of the presentation used in the consultations.¹⁰ It is important to note that these consultations were not conducted in English, and the statements and presentation were translated for this paper. Additionally, the consultations did not include a complete analysis of the GAMs but rather aimed to gauge the initial reactions of medical experts to the shape plots and their expected value in a practical context.

¹⁰ For the translated version of the presentation see doi.org/10.17605/OSF.IO/2WP6F.

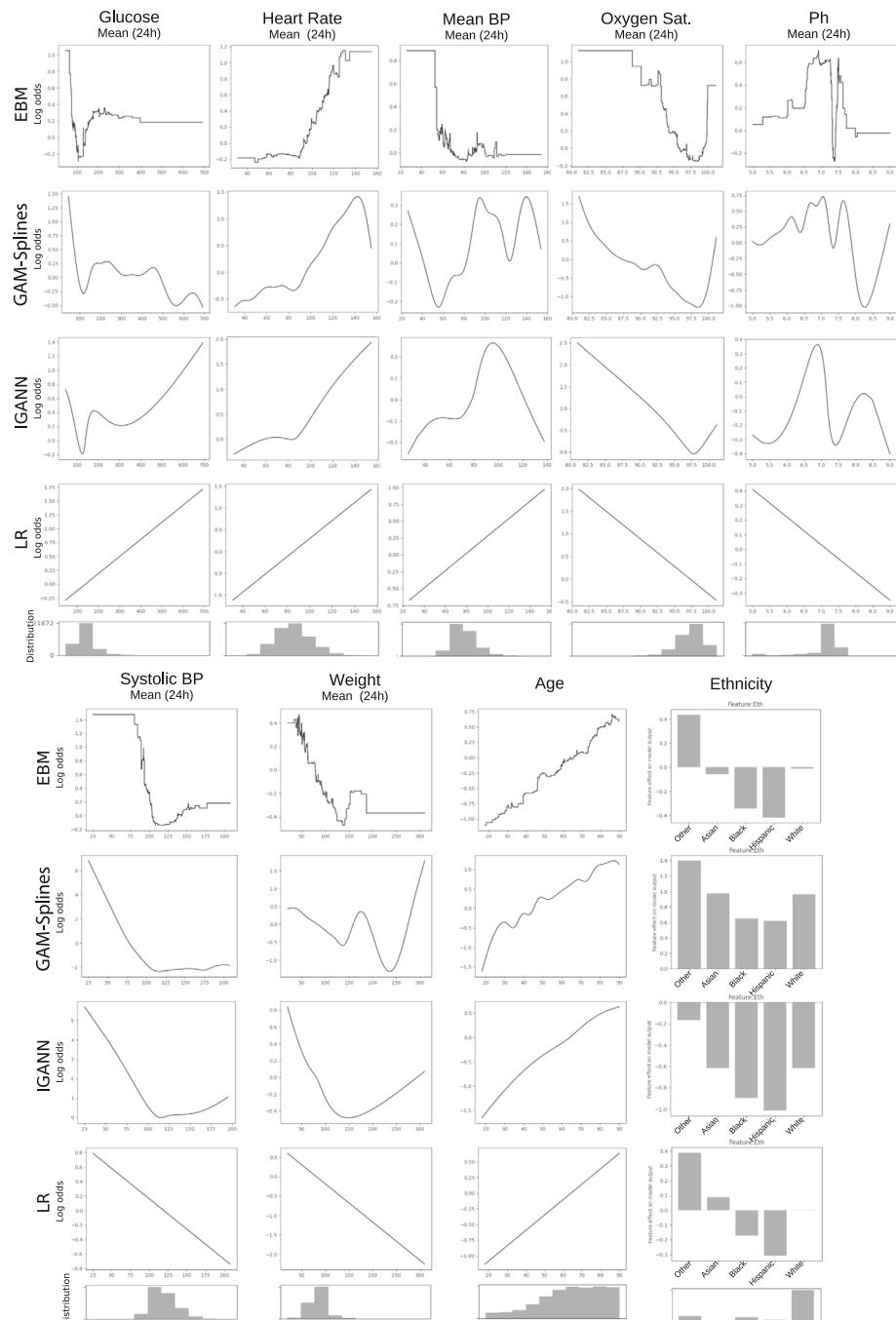


Fig. 7 Effect of 9 (out of 14) features on patient mortality, completing the shape plots presented in Fig. 4. Models were trained on the manually reduced dataset (D_{11-Man})

Table 12 Summary of medical experience

ID	Current position	Experience with ICU patients
ME1	Resident physician, internal medicine	<ul style="list-style-type: none"> - Responsible for ICU patients during transports - 8 years as paramedic
ME2	Resident surgeon, pediatric surgery	<ul style="list-style-type: none"> - Works in neonatal ICU - Provides intensive surgical care to premature infants - Daily pediatric ICU care - 5.5 years as pediatric surgeon at university hospital
ME3	Chief physician at anesthetic ICU	<ul style="list-style-type: none"> - Focused on ICU patients for 13 years - Responsible for 30-bed anesthetic ICU - 18 years as anesthesiologist
ME4	Physician in private medical office	<ul style="list-style-type: none"> - Supervised different ICUs (15 years in total) - 30 years as anesthesiologist

Appendix H: Direct quotes from medical experts

In this section, we present some direct quotes from the interviews with the medical experts that explain how we arrived at our conclusions from the consultations with them. These quotes allow the reader to get an unfiltered impression from the medical experts' perspective.

- The different opinions of the medical experts on the applicability of the shape plots of the different GAM models in the context of intensive care medicine.

ME1: "Well, I mean, this unsteadiness in the [EBM] graph perhaps suggests an accuracy that is ultimately not there after all. [...] That's why I wouldn't mind a flatter, slightly averaged graph."

ME2: "The important thing is the trend [...] not whether there is a slight reduction in [...] mortality risk [at a age of] 35 or whatever. [...] as a doctor working clinically at the bedside of an intensive care unit, I would think, no, I don't need it."

ME3: "For things like temperature and pH, where I have a relatively narrow range, I think a more precise representation [refers to the EBM] is perhaps more practical."

- A medical expert's perplexity regarding the increased mortality probability for patients with a GCST of 14.

ME1: "I do not quite understand the outlier at 14."

- Comments on the shape plots of blood pressure and temperature.

ME2: "So low blood pressure is always worse than very high blood pressure. Unless you get complications from high blood pressure, [...] like a cerebral hemorrhage."

ME1: "Especially in the lower temperature ranges, the middle graph [GAM-Splines] has a peak at about 35 degrees, which I can't quite work out where it comes from. Yes, as I said, I already know the one on the left [EBM]. I don't quite understand why the plateau is in the lower temperature ranges. From my point of view, I think the one on the right [IGANN] is the most plausible."

- Quotes on the difficulty of interpreting a shape plot that is based on the standard deviation of a feature rather than the mean of that feature.

ME4: "I find that really difficult. Yes, so the first thing that is irritating is that the Glasgow Coma Scale [...], that you don't start with 15, but as a clinician you tend to look at whether it's 3 or [...] 15, so at first I would think, well, I find that difficult and I would have to ask again, what do you mean by variability in the variable."

- Ideas on how to explain the negative monotonic trend between an increased variability of GCST and a lower prediction of mortality risk.

ME1: "The case with a high variability, which would be protective, is that a patient comes to the ICU after an operation, is still ventilated, then perhaps a GCST of three is entered on admission. And then, of course, [the patient] will be allowed to wake up as quickly as possible in the next hours. If the condition allows it. Then, of course, he may show a strong improvement. Of course, this is also a case that occurs frequently in ICUs, i.e., several times a day, where patients come in and wake up after an operation. I could just imagine that this is the reason for the system to see this high variability as protective."

ME3: "So the more common case is, of course, that I admit a patient from the OR who is still under anesthesia, has a GCS of 3, then I wake him up and he has a GCS of 15, so that's why the trend is simply more frequent."

- Quotes from medical experts expressing concern about incorporating the learned relationship between GCST standard deviation and mortality into a potential ICU decision support system.

ME2: "I find it difficult because it can go both ways. So it could be that you come up with a scale of 3 and then it's 15 or the other way around."

ME3: "There is also the case that I admit a patient with a GCS of 15 and now he has a cerebral hemorrhage and then 3 hours later has a GCS of 3."

- Ethical implications of ICU mortality prediction that emerged during the consultations.

ME3: "There is definitely an ethical component behind this, which is not insignificant in intensive care medicine [...] what do I do with this percentage [risk of death]? [...] the algorithm tells me that the patient has a 10% probability of survival [...], what is the conclusion then? Is the conclusion: he is going to die with a chance of 90% anyway, so we stop the therapy. Or do we say: He has a 10% chance, so let's try everything we can, to improve this 10% perhaps by taking some measures [...] so for the mortality risk, the doctor providing the therapy is not the ideal target group."

ME4: "So there are situations where you say: is it still worth it? To put it bluntly. When they're just totally sick and you say you'll try anything [...] then you can possibly prolong suffering. [describes example case in detail] And right at the beginning, if you had taken all these parameters into account, you may wouldn't have had to do that."

- Comments on the importance of integrating interaction terms into predictive models.

ME4: "Physiological systems always interact and the variables are all interdependent. So there is actually no single variable that is not connected to anything else, and to that extent, the interaction is important."

Author Contributions L.B., J.R., M.K., and P.Z. designed the study. L.B. and J.R. wrote the manuscript. In this process, M.K. and P.Z. provided detailed feedback and were involved in designing the overall structure.

L.B. and J.R. extracted and analyzed the data. All authors critically reviewed the manuscript and contributed to the interpretation of the results. L.B. and J.R. are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final draft of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. J.R., P.Z., and M.K. acknowledge funding from the Federal Ministry of Education and Research (BMBF) on "White-Box-AI" (Grant 01IS22080). L.B. and P.Z. acknowledge funding from the Federal Ministry of Education and Research (BMBF) on "AddICron" (Grant 16SV8995). This work was supported by an Academic Hardware Grant provided by Nvidia.

Data availability statement This study utilizes the MIMIC-III dataset, a publicly available and de-identified health-related database containing comprehensive information on ICU patients from Beth Israel Deaconess Medical Center (2001–2012). Access to the dataset requires completion of the CITI "Data or Specimens Only Research" course and a signed data use agreement. For instructions on obtaining access to the MIMIC-III dataset, please visit the official website.

Relevant links: MIMIC-III: <https://doi.org/10.13026/C2XW26> Project website: <https://mimic.physionet.org>. Access instructions: <https://mimic.mit.edu/gettingstarted/access/>.

Code Availability As described above, we based our feature extraction on the MIMIC-III benchmark developed by Harutyunyan et al. (2019). We forked their original Github repository and modified it to suit our needs.

* Our fork is available here: https://github.com/HB-Dynamite/mimic3-benchmarks_AoOR_data_export.

* For our experiments, we created a separate repository, available here: https://github.com/HB-Dynamite/Interpretable_ICU_predictions.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Agarwal, N., & Das, S. (2020). Interpretable machine learning tools: A survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 1528–1534.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. In *Advances in neural information processing systems* (vol. 34, pp. 4699–4711).

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, 373(6552), 284–286.

Bai, J., Fügener, A., Schoenfelder, J., & Brunner, J. O. (2018). Operations research in intensive care unit management: A literature review. *Health Care Management Science*, 21(1), 1–24.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123–1131.

Bertsimas, D., Pauphilet, J., Stevens, J., & Tandon, M. (2021). Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*.

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, pp. 25–60.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. London: Routledge.

Brunton, S. L., & Kutz, J. N. (2019). *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge: Cambridge University Press.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721–1730). New York, NY.

Chang, C-H., Tan, S., Lengerich, B., Goldenberg, A., Caruana, R. (2021). How interpretable and trustworthy are GAMs? In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 95–105). New York, NY.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY.

Coussemot, K., & Benoit, D. F. (2021). Interpretable data science for decision making. *Decision Support Systems*, 150, 113664.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on machine learning* (pp. 233–240). New York, NY.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint [arXiv:1702.08608](https://arxiv.org/abs/1702.08608)

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77.

Gerke, S., Minssen, T., Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial intelligence in healthcare* (pp. 295–336). Elsevier.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a “Right to Explanation”. *AI Magazine*, 38(3), 50–57.

Gunning, D., & Aha, D. (2019). DARPA’s explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44–58.

Halpern, N. A., & Pastores, S. M. (2010). Critical care medicine in the united states 2000–2005: An analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical Care Medicine*, 38(1), 65–71.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Steeg, G. V., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), 96.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398).

Hegselmann, S., Volkert, T., Ohlenburg, H., Gottschalk, A., Dugas, M., Ertmer, C. (2020). An evaluation of the doctor-interpretability of generalized additive models with interactions. F. Doshi-Velez et al. (Eds.), *Proceedings of the 5th machine learning for healthcare conference* (vol. 126, pp. 46–79). PMLR.

Hollinger, A., Gayat, E., Féliot, E., Paugam-Burtz, C., & Fournier, M-C., Duranteau, J. others., (2019). Gender and survival of critically ill patients: Results from the frog-icu study. *Annals of Intensive Care*, 9, 1–8.

Huang, G-B., Zhu, Q-Y., Siew, C-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3), 489–501.

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbisch, T., Esteban, C., & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3), 364–373.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning* (vol. 112). Springer.

Johnson, A., Pollard, T., Mark, R. (2016). *MIMIC-III clinical database*. PhysioNet.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1), 1–9.

Johnson, M., Albizri, A., & Simsek, S. (2022). Artificial intelligence in healthcare operations to enhance treatment outcomes: A framework to predict lung cancer prognosis. *Annals of Operations Research*, 308(1), 275–305.

Kocheturov, A., Pardalos, P. M., & Karakitsiou, A. (2019). Massive datasets and machine learning for computational biomedicine: trends and challenges. *Annals of Operations Research*, 276(1), 5–34.

Kramer, A. A., Dasta, J. F., & Kane-Gill, S. L. (2017). The impact of mortality on total costs within the ICU. *Critical Care Medicine*, 45(9), 1457–1463.

Kraus, M., Feuerriegel, S., & Saar-Tsechansky, M. (2024a). Data-driven allocation of preventive care with application to diabetes mellitus type ii. *Manufacturing & Service Operations Management*, 26(1), 137–153.

Kraus, M., Tscherntutter, D., Weinzierl, S., & Zschech, P. (2024b). Interpretable generalized additive neural networks. *European Journal of Operational Research*, 317(2), 303–316.

Kundu, S. (2021). AI in medicine must be explainable. *Nature Medicine*, 27(8), 1328–1328.

Lengerich, B., Tan, S., Chang, C- H., Hooker, G., Caruana, R. (2020). Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In S. Chiappa and R. Calandra (Eds.), *Proceedings of the twenty third international conference on artificial intelligence and statistics* (Vol. 108, pp. 2402–2412). PMLR.

Lou, Y., Caruana, R., Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining-KDD '12* (p.150). Beijing, China.

Lou, Y., Caruana, R., Gehrke, J., Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining-KDD '13* (pp. 623–631). New York, NY.

Malik, M. M., Abdallah, S., & Ala’raj, M. (2018). Data mining and predictive analytics applications for the delivery of healthcare services: A systematic literature review. *Annals of Operations Research*, 270(1–2), 287–312.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.

Moreno, R. P., Metnitz, P. G. H., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., SAPS 3 Investigators (2005). SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine*, 31(10), 1345–1355.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 807–814).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (General Data Protection Regulation)*.

Parliament and Council of the European Union (2021). *Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11), 1119–1125.

Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83, 112–134.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.

Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1), 3923.

Roncarolo, F., Boivin, A., & Denis, J- L., Hébert, R., Lehoux, P. (2017). What do we know about the needs and challenges of health systems? A scoping review of the international literature. *BMC Health Services Research*, 17, 1–18.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Saadatmand, S., Salimifard, K., Mohammadi, R., Kuiper, A., Marzban, M., Farhadi, A. (2022). Using machine learning in prediction of icu admission, mortality, and length of stay in the early stage of admission of covid-19 patients. *Annals of Operations Research*, pp. 1–29.

Sivaraman, V., Bukowski, L.A., Levin, J., Kahn, J.M., Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–18).

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5), e1379.

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *The Lancet*, 304(7872), 81–84.

Topuz, K., Uner, H., Oztekin, A., & Yildirim, M. B. (2018). Predicting pediatric clinic no-shows: A decision analytic framework using elastic net and Bayesian belief network. *Annals of Operations Research*, 263(1), 479–499.

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383(9), 874–882.

Wang, S., McDermott, M.B.A., Chauhan, G., Ghassemi, M., Hughes, M.C., Naumann, T. (2020). MIMIC-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM conference on health, inference, and learning* (pp. 222–235). New York, NY.

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18), 1837–1847.

Yang, C. C. (2022). Explainable artificial intelligence for predictive modeling in healthcare. *Journal of Healthcare Informatics Research*, 6(2), 228–239.

Yang, Z., Zhang, A., & Sudjianto, A. (2021). GAMI-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120, 180–192.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5, 1205–1224.

Zschech, P., Weinzierl, S., Hambauer, N., Zilker, S., Kraus, M. (2022). GAM(e) change or not? An evaluation of interpretable machine learning models based on additive model constraints. In *Proceedings of the 30th European conference on information systems (ECIS)*. Timisoara, Romania.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.