Machine Learning Algorithms for Early Sepsis Diagnosis: Predictive Power, Explainability and Economic Advantage

Jan Frąckowiak, Marcin Chlebus[1]
jj.frackowiak2@student.uw.edu.pl, m.chlebus@uw.edu.pl
Department of Data Science, Faculty of Economic Sciences, University of Warsaw[1]

Abstract

The adoption of predictive algorithms in sepsis diagnosis promises to enhance diagnostic accuracy, enable earlier detection, and achieve significant cost savings; however, trust in such systems remains limited within the medical community. In response to this problem and the associated obstacles, the article investigates the efficacy of neural networks and gradient boosted trees (GBT) paired with eXplainable Artificial Intelligence (XAI) techniques. Models' results are validated against requirements for solutions adoptable in clinical settings and compared with recent literature. Contrary to initial suggestions based on leading voices in the field, GBT emerge as a preferred diagnostic tool, demonstrating predictive power comparable to neural nets and superior economic advantage. Verification of the use of XAI techniques reveals improved comprehension of model outcomes for healthcare practitioners, while economic analysis underscores the cost-effectiveness and potential of adopting similar solution in clinical practice.

Keywords: Sepsis, Machine Learning, Temporal Neural Networks, Gradient Boosted Trees, XAI

1. Introduction

Sepsis, a critical medical condition triggered by the body's extreme response to infection, poses a significant challenge in healthcare due to its rapid progression and high mortality rates (WHO, 2024). Timely identification and intervention are paramount to improving patient outcomes (Kim and Park, 2019). In recent years, the integration of artificial intelligence (AI) technologies has emerged as a promising approach to enhance early prediction and management of sepsis (Yang et al., 2023). Similar to cardiac arrest, where early intervention is crucial, the utilization of AI holds potential to revolutionise sepsis care by enabling healthcare providers to anticipate and address the condition before it escalates.

However, despite the general precision and explanatory power of AI models, the practical implementation of such algorithms remains a considerable challenge. This challenge is exacerbated by the novelty of AI solutions, which may not immediately garner trust from naturally conservative communities such as lawyers, banking industry or healthcare practitioners. To address this, requirements for predictive algorithms that could help win the trust of such communities should be specified.

Fan et al., 2020 distinguish four core domains to support adoption of AI-DDS (AI Diagnostic Decision Support) depicted on Figure 1. "Reasons to use" domain divides to ability of a tool to address a pressing clinical need and improve patient care and outcomes (alignment) and the tool's affordability both to the patient and health system, including the incentives for the provider, patient, and health system (incentives). The remaining three domains encompass "Means to use", which de-

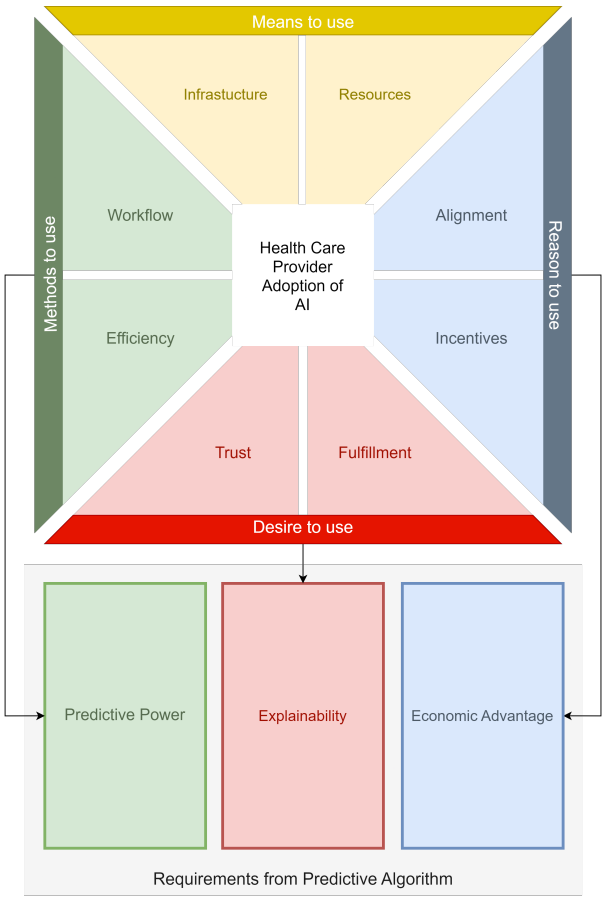


Figure 1: Core domains to support adoption of AI-DDS and implied requirements from predictive algorithm.
Source: Own preparation basing on graph from Fan et al., 2020 article.

notes the technical feasibility and capability to implement a tool effectively, "Desire to use", which reflects the trust and acceptance of the tool within the community of end-users and "Methods to use", which pertains to the quality of methods employed and their seamless integration into existing workflows.

Drawing from characteristics emphasised by regulations on the field of utilising software as medical devices (such as the severity of clinical trials and stability of model outcomes) and the requirements specified in aforementioned review, we could distinguish four characteristics of machine learning algorithm, which emerge as particularly significant in terms of solution adoptability:

Predictive Power: The algorithm must demonstrate high accuracy and reliability in predicting disease onset or progression. Rigorous testing and validation against real-world data sets should be conducted to ensure its effectiveness across diverse patient populations and clinical settings. Related to "Methods to Use" domain.

Ease of Technical Implementation: The algorithm should be seamlessly integrated into existing healthcare systems and workflows without causing disruptions. Compatibility with Electronic Health Records (EHRs) and interoperability with other medical devices and software are essential for smooth adoption by healthcare professionals. Related to "Means to Use" domain.

Economic Advantage: Highlighting the economic benefits and business advantages gained by implementing the AI application can incentivise stakeholders, including healthcare institutions and insurance providers, to embrace the technology. This may include cost savings through early intervention, reduced hospital re-admissions, and improved resource allocation. Related to "Reason to Use" domain.

Explainability: Ensuring the transparency and interpretability of the AI model is crucial for building trust among clinicians and other stakeholders. The algorithm should provide clear and understandable rationales for its predictions, allowing healthcare providers to comprehend the underlying factors influencing its decisions and enabling them to validate its recommendations against their clinical judgment. Related to "Desire to Use" domain.

By addressing these key features, AI-based predictive algorithm supporting diagnosis could overcome barriers to adoption and gain the trust of conservative communities in healthcare and beyond.

2. Literature Review

Yang et al., 2023 conducted a survey of the state of AI application in systems supporting sepsis diagnosis. They found that conventional assessment tools like the quick sequential organ failure assessment (qSOFA) often underperform in the diagnosis and treatment of sepsis, as patients present with diverse clinical manifestations and respond differently to treatments. This necessitated more precise and personalised treatment option, which can be offered by AI models thus improving patient outcomes.

There are numerous cases of successful developments of predictive models related to sepsis. For instance Escobar et al., 2020 have employed logistic regression, SAPS-II scoring prediction, and XGBoost algorithms to construct three distinct machine learning models. These models were designed to anticipate the 30-day mortality rate among sepsis patients in the MIMIC-III database, with the XGBoost model demonstrating the highest accuracy, reaching 85%.

Mao et al., 2018 build a case for the wider use of machine learning-based sepsis diagnosis. The authors used only six vital signs (heart rate, respiratory rate, peripheral capillary oxygen saturation, temperature, systolic, and diastolic blood pressure) to train a Gradient Boosting Tree (GBT) model predicting sepsis up to 48 hours in advance of onset, achieving an AUC-ROC of 0.83.

Certain models have already been applied in clinical settings. The Neonatal early-onset sepsis (EOS) calculator is founded on a predictive risk model developed through case-control design involving 608,014 newborns born in 14 U.S. hospitals. This model underwent further refinement using logistic regression. The EOS Calculator (http://kp.org/eoscalc) estimates the risk of EOS based on five objective maternal risk factors and four clinical neonatal risk factors. (Zayek et al., 2020)

Furthermore, researchers have employed statistical and machine learning approaches to develop a model known as PEDSEPS-GBM (Ying et al., 2021). They have also created a web calculator for this model to assist clinicians in early detection of sepsis in children.

Yang et al., 2023 observe that despite its indisputable potential in medical applications, current practical research on AI, needs to cope with challenges such as insufficient applicability and limited alignment with traditional diagnostic logic. These constraints significantly impede the exploration of AI's clinical utility, and call for methods with predictable and comprehensible results.

The main inspiration for this work was the 2020 article where Lauritsen et al., 2020 propose a system of explainable AI early warning score (xAI-EWS) which promises the integration of XAI in clinical settings and possibility of wider adoption of diagnostic algorithms based on deep learning. One key advantage highlighted is the enhanced trust and transparency offered by presented xAI model, that provides clinicians with clear insights into how predictions are made, fostering a better understanding and trust in AI-generated results. This transparency not only boosts confidence in the system, but should also help clinicians understand the data driving predictions. Additionally, the explainable nature of the model ensures compliance with regulations like the EU GDPR and the US FDA.

As the authors of the publication argue, implementation of XAI methods may allow for replacing traditional EWS (Early Warning Score) and SOFA (Sequential Organ Failure Assessment) diagnostic system with an implementation of a deep learning algorithm which otherwise would be considered a black-box.

The proposed system is composed of a TCN (Temporal Convolutional Neural Network) which serves as a predictive algorithm outputting values from 0.0 to 1.0 and DTD (Deep Taylor

Decomposition) as the explanation module accounting for the temporal nature of the input data. The hope of 3-layered TCN performing better than traditional Recurrent Neural Networks architectures was based on the belief in dilated convolutions increasing the receptive field of higher-order neurons to entire sequence of input data, as well as performance of CNN's in computer vision tasks. Adoption of TCN also allowed for a significant decrease of training time (30 minutes to convergence). The utilised window of observations was 24 hours of 33 clinical measurements as input for the model. These settings were used to predict the probability of future sepsis, acute kidney injury (AKI) and acute lung injury (ALI) during admission.

The development and implementation of the xAI-EWS model for sepsis prediction came with a set of inherent challenges and limitations, and nonetheless authors report a great increase of AUC-ROC in comparison to (modified) EWS, Gradient-Boosting vital signs assessment method and SOFA during the frame of 24 hours before onset.

One notable hurdle reported by the authors is the imbalanced classification problem stemming from the low prevalence rates of sepsis, acute kidney injury (AKI), and acute lung injury (ALI) with rates as low as 2.44%, 0.75%, and 1.68% respectively. In such conditions training the model effectively becomes challenging due to the disproportionate representation of classes. In attempts to address this issue, the study employed oversampling techniques on the positive class, yet this didn't significantly enhance model performance and stretched output probabilities into a wider range. The aspect of interpretability of predictions was addressed by simple visual explanations from DTD module displaying both global and local importance of input variables. DTD itself is a method of layer-wise relevance propagation which allows for calculating relevance score for each of the input features in a way similar to backpropagating an error. The authors estimated the back-propagated relevance only on the patients with positive classes.

The study highlights the importance of refining ground truth definitions for critical illnesses like AKI and ALI, as well as validating the model's performance across different populations to ensure its ability to generalise. Compliance with regulations such as the EU GDPR and FDA further adds complexity to the development and implementation process. As the xAI-EWS currently relies on a limited set of clinical parameters, expanding its scope to incorporate a broader range of features presents an ongoing challenge. Authors conclude that despite these obstacles, the iterative model development process, coupled with continuous collaboration with clinical experts, offers avenues for addressing these limitations and advancing the application of xAI in sepsis prediction.

Yet another major inspiration for this work as well as the source of data was the 2019 Physionet challange devoted to creating an algorithm of early sepsis diagnosis. The contest result was described by Reyna et al., 2020.

Using data from three separate hospital systems for the challenge brought the diversity and real-world applicability to the evaluation of algorithms. The challenge organisers aimed to assess the generalisability and robustness of the algorithms.

The algorithms submitted to the challenge underwent evaluation based on a utility score framework. This score penalised algorithms predicting sepsis earlier than 12 hours before onset while rewarding those predicting sepsis within the 12 to 3-hour window. Ultimately, the algorithm achieving the highest utility score on the full sequestered dataset from hospital systems A, B, and C emerged as the challenge winner.

The team securing the top position (Morrill et al., 2019) leveraged the temporal relationships between features by employing signature transforms. These transforms, involving iterated integrals of the time series data, captured interactions and dependencies within the temporal components, thereby assessing the geometric structure or shape of the data path. On the other hand, the second team's approach (Reyna et al., 2020) introduced a way to address missing data problem by incorporating flags indicating missing values alongside vital signs and their variances. The third team (Zabihi et al., 2019) opted to include mean and variance of sequence lengths as well as variances of selected vital signs. The performance of these algorithms was tracked using the area under the ROC curve (AUC), yielding scores of 86.8%, 86.3%, and 83.3%, respectively.

Despite the diverse feature engineering strategies employed, all three top-performing teams relied on variants of gradient boosted decision trees. This fact underscores the adoptability and effectiveness of this class of algorithms in the task of sepsis diagnosis. However, it also prompts a natural question regarding the viability of TCN-based solution proposed by authors of the XAI-EWS system, especially considering its reported superiority over well-established GB-vital diagnostic systems. The choice of TCN over RNN architectures, which are standard in sequence processing, is also controversial. Physionet dataset (Reyna et al., 2020), provides an opportunity to validate discussed variants of neural net models performance against each other and GBT models, which demonstrated strong performance in the challenge.

To gain the trust of the medical community, it is essential that the influence of variables on predictive models aligns with established medical knowledge and meets performance standards. While establishing medically supported expectations for sepsis diagnostic models presents a challenge, numerous studies exploring sepsis risk drivers have been conducted and can be used to inform and shape such expectations effectively.

According to Li et al., 2022, main sepsis risk drivers include age, D-dimer, albumin, creatinine, and prothrombin time. These variables reflect underlying health conditions that can predispose an individual to sepsis. Furthermore, a recent study by Aygun et al., 2024 observed that patients with sepsis had significantly higher levels of indicators related to increased metabolism and immune response. These indicators included respiratory rate, heart rate, body temperature, leukocyte particle concentration, C-reactive protein, procalcitonin, neutrophil–lymphocyte count ratio, and plaque. Conversely, the sepsis group had substantially lower levels of hemoglobin, oxygen saturation, and systolic blood pressure.

Variables indicating an increased metabolic state and the body's response to illness, such as increased heart and respiratory rates, have been shown to correlate with higher sepsis risk Shashikumar et al., 2017, Lee et al., 2021. Similarly, an el-

evated leukocyte count is associated with sepsis Rimmer et al., 2022. Factors indicating liver or kidney dysfunction, which are often impaired due to sepsis, also may be indicative of its risk.

Increased body temperature, which might naively be considered a clear indicator of sepsis, actually shows a more complex relationship with the illness. While fever is often reported in cases of mild or moderate sepsis, lower body temperatures also can be indicative of a more severe progression of the disease Rumbus and Garami, 2019.

Factors describing good blood health and overall bodily function generally decrease the risk of sepsis. Such factors include high respiratory efficiency Qu et al., 2023, and adequate systolic/diastolic blood pressure Gao et al., 2023. Additionally, hemoglobin and hematocrit levels Agnello et al., 2021, as well as platelet counts Hua et al., 2023, are inversely related to sepsis risk.

In the aforementioned study (Aygun et al., 2024), low oxygen saturation was also identified as a symptom of sepsis, a view widely supported in the medical literature (Avendaño-Ortiz et al., 2018). However, as per Textoris et al., 2011, hyperoxemia may also be linked to increased sepsis severity, suggesting that this variable does not necessarily have a linear effect on sepsis risk

Demographic factors like gender and age are consistently reported as significant indicators of sepsis risk. Lakbar et al., 2023 found that sepsis and septic shock are more prevalent in men than in women. A systematic review by Fathi et al., 2019 involving 2,978 articles confirmed that older age and male gender are associated with an increased risk of sepsis.

The length of stay in the ICU is another highly informative feature. It often reflects the mortality among patients, general long term prognosis (Rodrigues et al., 2024) and probability of septic shock (Tuttle et al., 2023). Cardoso et al., 2011 calculated that each hour of waiting for ICU admission was independently associated with a 1.5% increased risk of ICU death, highlighting the critical importance of timely diagnosis and intervention.

Given the rapid pace of healthcare AI development, we now see an increasing number of studies and publications dedicated specifically to the economic impact and utility of these technologies. Khanna et al., 2022, recognised as the first comprehensive study in the field of AI economics, assessed the impact of AI on healthcare costs. Their findings demonstrated that AI significantly lowers healthcare costs compared to traditional methods. The authors highlighted the role of time efficiency in diagnosis, where AI algorithms exhibit superior performance, especially as the number of patients increases. This effect led them to estimate cost savings in diagnosis of USD 1,666.66 per day per hospital during the first year of simulation on the studied sample.

The cost impact of early detection and treatment of sepsis was also addressed by Ericson et al. in their 2022 study. This work presented a decision tree model that incorporates the conditions necessary for AI implementation and support in the context of sepsis diagnosis. Authors assumed that a machine learning algorithm capable of detecting sepsis only three hours before clinical diagnosis could reduce the cost per ICU patient by 0.5%. Specifically, they found that the total cost per patient using such an algorithm would be €16,436, compared to €16,512 for current practices, resulting in a potential cost saving of €76 per patient. When applied to the Swedish healthcare system, this would translate to an aggregated yearly cost saving of €2,798,915.

The most significant cost savings were attributed to a shorter average length of stay in the ICU, with a reduction of 0.16 days (8.9%) per patient when using an algorithm like NAVOY® Sepsis, compared to current practices (€10,322 vs. €11,331 per patient for ICU hospitalization). The studies also underscored the distinguishing effects of true positive and false negative outcomes, regarding the timing of treatment. As discussed by the authors, in the model's base case, true positives identified by the sepsis prediction algorithm were assumed to receive treatment three hours earlier than true positives under current practices. Conversely, for false negatives, appropriate treatment was assumed to be delayed by an additional three hours compared to the timing in current practices.

3. Aim and Hypotheses

The following section outlines three main objectives of the study. These aims address the requirements for medical diagnostic tool (described in introduction) and respond to recent literature in both the medical field and AI systems in healthcare (described in literature review). Each of these aims was expressed in hypotheses that are validated during the study.

3.1. Verifying trained models against the requirements for AI-DDS adoptability

The primary aim of this work is to use the example of early diagnosis of sepsis to examine how three key characteristics of adoptable diagnostic system - predictive power, explainability and economic advantage - can be ensured during machine learning model development.

Drawing on insights from literature and proprietary experiments, a set of algorithms was trained and evaluated against these requirements.

The fourth aspect, ease of implementation, is not discussed as it falls outside the scope of this research and is more related to infrastructure capabilities. However, with the availability of open-source methods, implementing AI does not require using supercomputers.

To address the requirement for predictive power, the article examines the process of model selection, achieved performance metrics, and the dataset used for experiments. Various machine learning algorithms are evaluated to identify the most suitable one for predicting sepsis onset and progression. Additionally, the dataset's quality and representativeness are assessed to ensure the reliability of the results.

Explainability plays a crucial role in bridging the gap between powerful, yet opaque, algorithms and the trust of medical practitioners. The article tackles this challenge by exploring techniques and strategies to enhance the AI model's explainability, such as Shapley values, feature importance or partial

dependence profiles. Produced explanations are compared to expectations formed based on medical literature and other studies to validate the models against established knowledge.

The economic viability of the trained algorithms is evaluated by estimating potential cost savings compared to traditional clinical diagnosis. This includes examining the financial benefits of early intervention, reduced hospital stays, and improved patient outcomes, highlighting the potential for significant cost savings and increased efficiency in healthcare settings.

By systematically addressing predictive power, explainability and economic advantage, this article aims to provide valuable insights and practical guidance for utilisation of AI technologies in the early diagnosis and management of sepsis. This aim was associated with the following hypothesis:

I. Available open-source algorithms, with appropriate preparation, can meet the requirements for AI-DDS (Artificial Intelligence-Driven Decision Support).

3.2. Validation of TCN utilisation in the task of sepsis diagnosis

The secondary aim is to validate the application of Temporal Convolutional Network (TCN) which was proposed as the most performing algorithm in the Explainable AI Early Warning Score (XAI-EWS) system (Lauritsen et al., 2020). Notably, the choice of TCN for the task of sepsis diagnosis is controversial since LSTM serves as the standard for sequence processing, and GBTs are widely used in AI systems for sepsis diagnosis. Validation was performed on a different dataset and by comparing TCN against more commonly used models to determine if a TCN-based system could maintain or improve predictive performance across diverse clinical data. This aim was associated with the following hypothesis:

II. TCN (Temporal Convolutional Network) will outperform other models in terms of predictive power.

3.3. Verifying trained models against medical knowledge

Based on medical literature, we can establish expectations regarding the model's interpretability in the context of predicting sepsis risk and perform a validity check on how this risk is assigned to patients. Specifically, two groups of variables, representing expected negative and positive effects on the model's output, are outlined in the following hypotheses:

III. ICULOS, age, male gender, and variables indicating increased metabolism or ongoing immune response of organism (respiratory rate, heart rate, increased leukocytes levels) will have a positive impact on model output.

IV. Variables defining good condition and efficiency of organism (hemoglobin/hematrocit, systolic and diastolic blood pressure, platelets count) will have a negative impact on model output.

These two hypotheses could be verified only with the help of XAI techniques.

3.4. Demonstrating approach for assessing the economic value of the solution.

The fourth aim of this article is to propose a satisfactory approach for assessing the economic value of most performing models. By analyzing cost savings using specific assumptions based on studies by Paoli et al., 2018 and Liang et al., 2020 the financial benefits of early detection and treatment are estimated. Selected cost estimation method accounts for the influence of predictions resulting in early or extended treatment.This aim was associated with the following hypothesis:

V. Applying developed algorithms to diagnose patients from the test sample would lead to cost savings.

4. Methods

This section details the modeling pipeline developed using data from the Physionet Challenge (Reyna et al., 2020), resulting in effective predictive models. It also covers the explainability measures implemented and introduces a proposed algorithm for evaluating economic advantage. Together, these elements provide an overview of the methods employed to address the three essential requirements of AI-DDS systems: Predictive Power, Explainability, and Economic Advantage.

4.1. Data

The dataset proposed by Physionet consisted of hourly measurements of 41 features including vital signs, laboratory values extracted from blood analysis as well as basic demographic information. The available dataset was divided into two subsets indicating respective ICU units: training set A (20,336 subjects) and B (20,000 subjects). Affiliation with hospital was reflected with binary variable - Unit1. Due to the lack of a natural way to infer affiliations, any missing values for Unit1 were imputed. The entire dataset contained a total of 813,712 observations. All variables utilised in modelling are described in Table 1.

For the purpose of modeling, feature engineering was performed to augment the number of variables considered in the analysis. Initially, medically justified interactions were computed for variables with less than 70% missing values. Age and respiration rate (Resp) were included for potential interaction effects, given that variations in respiratory function might be age-dependent. Concerning cardiovascular function, an interaction term between heart rate (HR) and systolic blood pressure (SBP) was derived. This interaction aimed to capture the relationship between heart rate and blood pressure, which can serve as an indicator of cardiovascular function. Further interactions included: in the context of hemodynamic stability, an interaction term involving mean arterial pressure (MAP) and diastolic blood pressure (DBP); for oxygenation assessment, an interaction term between oxygen saturation (O2Sat) and respiration rate (Resp); regarding temperature regulation, an interaction term between body temperature (Temp) and respiration rate (Resp) was computed. Changes in body temperature alongside respiration rate, in many cases serve as good proxy of

Variable	Description
Vital signs	
HR (10%)	Heart rate
O2Sat (13%)	Pulse oximetry
Temp (66%)	Temperature
SBP (15%)	Systolic BP
MAP (12%)	Mean arterial pressure
DBP (31%)	Diastolic BP
Resp (15%)	Respiration rate
EtCO2 (96%)	End tidal CO2
Laboratory values	
Platelets (94%)	Blood components
WBC (94%)	Blood components
Hgb (93%)	Blood components
Hct (91%)	Blood components
Calcium (94%)	Blood chemistry
Glucose (83%)	Blood chemistry
PTT (97%)	Blood clotting
Fibrinogen (99%)	Blood clotting
BaseExcess (95%)	Excess bicarbonate
HCO3 (96%)	Bicarbonate
FiO2 (92%)	Inspired oxygen
SaO2 (97%)	Oxygen saturation
pH (93%)	Arterial gas
PaCO2 (94%)	Arterial gas
BUN (93%)	Kidney function
Creatinine (94%)	Kidney function
Lactate (97%)	Acid level
AST (98%)	Liver, kidney function
Bilirubin total (96%)	Liver function
Bilirubin direct (100%)	Liver function
Potassium (91%)	Liver function
Magnesium (94%)	Electrolyte balance
Chloride (95%)	Electrolyte balance
Phosphate (96%)	Mineral metabolism
Alkaline Phosphatase (98%)	Enzyme activity
Troponin (99%)	Cardiac biomarker
Demographics	
Age (0%)	Patient info
Gender (0%)	Patient info
Unit1 (39%)	ICU identifier
Unit2 (39%)	ICU identifier
HospAdmTime (0%)	Hospital-ICU time
ICULOS (0%)	ICU stay length
Engineered Features	
hr_o2sat_interaction	HR * O2Sat
hr_sbp_interaction	HR * SBP
mpa_dbp_interaction	MAP * DBP
o2sat_resp_interaction	O2Sat
temp_resp_interaction	Temp*Resp
age_resp_interaction	Age*Resp
First differences calculated on:	Temp, DBP, Resp, SBP, O2Sat, MAP, HR (< 70%)
Target with distribution	
SepsisLabel (0%)	Negative Patients: 92.7%, Positive Patients: 7.3%

Table 1: Variables utilised in training. Information about the percentage of missing values in brackets (%).
Source: Own preparation.

the body's response to infection and other physiological stressors. These interactions together with first differences of selected variables were integrated into the modeling framework to capture the impact of more complex physiological relationships on clinical outcomes.

4.2. Modelling pipeline

General description:

Figure 2 depicts machine learning pipeline designed for preparing and validating models with respect to three key requirements for AI-DDS adoptability. The first part of the pipeline demonstrates how neural networks and gradient-boosted trees were trained. The process begins with patient data input, followed by subsequent steps: extending the SepsisLabel for positive patients, Bayesian Ridge imputation for missing data, calculating interactions and first differences, and applying Min-Max scaling for data standardization.

The preprocessed data was then utilised in two parallel model training streams: one focused on neural networks, including ANN, LSTM, and TCN models, and the other on gradient-boosted trees, comprising XGBoost, LightGBM, and CatBoost models. Both streams underwent hyperparameter tuning through AUC-ROC maximization using grid search, with neural networks trained on 20-row moving windows and gradient boosted trees employing 5-fold cross-validation.

Following model training, a final comparison selected the best-in-class models, which were further evaluated for economic advantage and explainability. The economic assessment evaluated potential clinical benefits by quantifying time savings in early sepsis detection. The explainability assessment utilised SHAP and Dalex packages to offer deeper insights into the model's predictions.

Development:

The two main challenges faced during data preparation were the problem of missingness and severe class imbalance for the target variable. First of these problems was tackled with imputation with BayesianRidge algorithm. Scikit-Learn (Pedregosa et al., 2011) Python library supports numerous strategies of multivariate imputation, including Bayesian-Ridge Regression ranking best in comparisons found in literature (González-Vidal et al., 2020, M. Mostafa et al., 2020).

Thanks to imputation it became possible to estimate model on entire set of features and afterwards compare their importance. With the sample size of 569,909 observations in training set the curse of dimensionality was not a significant concern when estimating models on full set of features. Neural Net's are also believed to be robust to the curse of dimensionality (Poggio and Liao, 2018), while Gradient Boosted Models often serve as feature selectors (Saheed, 2023).

The second problem - class imbalance - is usually tackled with the use of a combination of undersampling and oversampling techniques such as SMOTE (Chawla et al., 2002. However changing the sample should be reconsidered in the perspective of utilising XAI methods which usually employ statistical methods of variables influence.

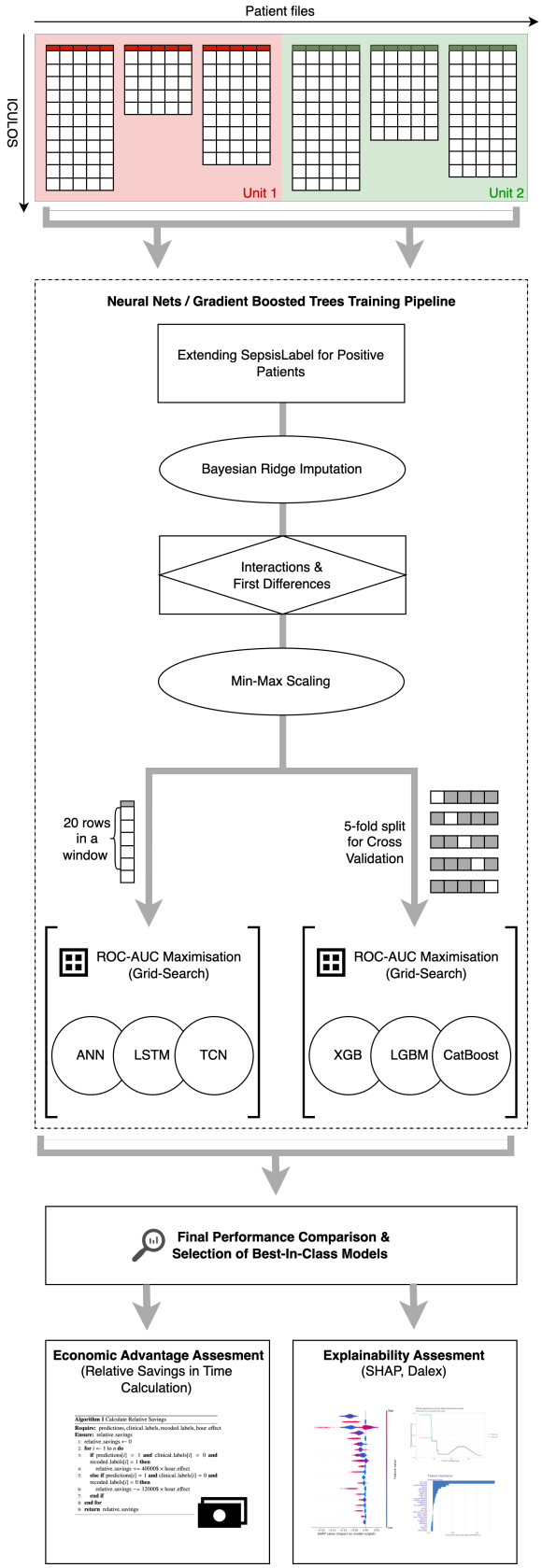


Figure 2: Steps of Models' Development and Assessment.
Source: Own preparation.

IQBAL and Sikdar, 2023 demonstrated instability of results of classifiers trained on synthetic data with the use of XAI, questioning the reliability of this method. Thus, in order to guarantee representativeness of the sample its original size was preserved.

Nevertheless, because the machine learning model was developed with the intention of classifying sepsis as early as possible, all sepsis labels for patients who had at least one positive clinical diagnosis of sepsis were recoded to one. This adjustment improved the class imbalance, resulting in 15.5% of the observations being positive cases.

After recoding the label and transformations (imputation, feature engineering and Min-Max scaling) the data was ingested to respective models with the aim of optimising area under receiver operating curve (AUC-ROC) expressing models' discriminative power.

Neural network architectures (TCN, LSTM, ANN) were trained with moving windows of 20 observations and a mini-batch size of 32. The loss function employed was the standard binary cross-entropy. To better handle class imbalance, *pos_weight* parameter was introduced, set to the inverse ratio of positive to negative examples in the training set (8), as recommended by PyTorch official documentation (PyTorch,). After pretraining, the best hyperparameters for each model class were selected to optimise AUC-ROC, evaluated simultaneously on both training and validation sets. For efficiency of grid-search median stopping rule was applied.

In case of models based on Gradient Boosted Trees AUC-ROC was maximised with the use of 5-fold cross validation performed on training set. Only after optimisation the metric was calculated on the validation and test set. Selected GB model took one observation of data for prediction.

15% of unique patients data was allocated to validation set and 15% to test set, while the remaining 70% was utilised during training. The data split was performed in a stratified manner, to preserve the proportion of patients with disease (15.5%). Apart from the usual training and test sets, a validation set was separated to provide an independent sample for unbiased optimization of the model's discriminative power and threshold tuning.

4.3. Architectures

This section describes architectures utilised during training. The selection of these algorithms is common among studies concerning sepsis detection and they are known for their performance.

Artificial Neural Networks (ANNs) (McCulloch and Pitts, 1943) are a fundamental type of neural network inspired by the structure and function of the human brain. ANNs consist of an input layer, one or more hidden layers, and an output layer. Each layer contains neurons (nodes) that process input data through weighted connections and non-linear activation functions. The process of adjusting these weights with respect to gradient of the loss function (backpropagation) during training enables the network to learn complex patterns and relationships within the data.

ANNs are highly flexible and can model complex non-linear relationships, making them suitable for a wide range of predictive tasks. In the context of sepsis prediction, ANNs can effectively handle large datasets with multiple features, such as vital signs, laboratory results, and patient demographics. However, standard ANNs do not inherently account for the temporal dynamics of patient data, which can be crucial for understanding the progression to sepsis. Thus this architecture was also applied to a moving window of observations.

Temporal Convolutional Networks (TCNs) (Lea et al., 2017) are designed to handle sequential data by applying convolutional neural network (CNN) principles. Unlike traditional CNNs used in image processing, TCNs use causal convolutions to ensure that predictions at any time step are influenced only by inputs from former time steps. This structure preserves the temporal order of the data.

TCNs are particularly efficient for processing sequences due to their ability to perform parallel computations of convolutions, which makes them faster in training than recurrent models like LSTMs. Additionally, TCNs can capture long-range dependencies through their hierarchical structure, where deeper layers aggregate information from increasingly larger temporal windows. This makes TCNs well-suited for time-series data in sepsis prediction, where understanding the evolution of clinical measurements over time can be crucial.

Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997) are a specialised type of recurrent neural network (RNN) designed to address the limitations of traditional RNNs in capturing long-term dependencies (e.g. the problem of exploding gradient). LSTMs incorporate memory cells and gating mechanisms (input, forget, and output gates) that regulate the flow of information, enabling the network to maintain and update information over extended time periods.

LSTMs are highly efficient at modeling sequential data with temporal dependencies, making them ideal for tasks where the order and timing of events are critical. Owing to hidden state transfer, LSTMs can capture how early signs and symptoms progress into sepsis by maintaining information across varying time lags. This ability make LSTM a suitable tool in predicting sepsis onset

CatBoost (Dorogush et al., 2018, is a model based on gradient boosted trees, including mechanisms to reduce overfitting. One of this mechanism is ordered boosting, which builds each tree using a random permutation of the dataset to ensure that the model does not rely too heavily on single observations.

XGBoost 2.0 (Extreme Gradient Boosting) (Chen and Guestrin, 2016) is an optimised gradient boosting library that enhances the performance and efficiency of traditional gradient boosting algorithms. XGBoost introduces several innovations, including a regularization to prevent overfitting, a tree-pruning algorithm, and the use of second-order gradient information to improve accuracy.

XGBoost is known for its speed and performance, making it a popular choice for many machine learning competitions and applications.

LightGBM (Light Gradient Boosting Machine) (Ke et al., 2017) is another gradient boosting framework that uses a histogram-based approach to speed up training and reduce memory usage. LightGBM constructs trees using a leaf-wise growth strategy with depth limitations, which leads to more efficient computation and better handling of large-scale data. LightGBM is particularly well-suited for large datasets with many features and observations.

All of the gradient boosting trees models are constructed as an ensemble of weak learners trained in a sequential manner. The basic mechanism involves iteratively adding models (typically decision trees) to correct the errors of the combined ensemble of previous models. Gradient boosting technique is popular because of its power, although it has a tendency to overfit, that is why all of the gradient boosting based models described above introduce some kind of regularization to ensure robustness of predictions.

In the pursuit of optimizing model performance, comprehensive grid searches were conducted for both neural network architectures, including LSTM, TCN, and ANN, as well as CatBoost, XGBoost, and LightGBM. These searches encompassed a broad array of hyperparameters, with each combination representing a potential model configuration. Here's a breakdown of the parameter spaces explored for each model:

For LSTM, the traversed parameter space included hidden sizes, number of layers, dropout rates, and learning rates, while employing ReLU as the activation function. Similarly, for TCN, configurations consisted of number of TCN channels, dropout rates, learning rates and number of layers. In the case of ANN, parameter exploration involved hidden sizes, dropout rates, and learning rates. The numbers of potential combinations for each neural net architecture were 12, 24, and 16, respectively for LSTM, TCN and ANN.

The search space for GBT models spanned over different hyperparameters. For CatBoost, this included iterations, learning rates, depths, L2 leaf regularization, bagging temperature, and border count. Similarly, XGBoost parameter grids were defined for learning rates, maximum depths, subsample ratios, column sampling by tree, gamma values, and minimum child weights. Finally, the optimization process for LightGBM included learning rates, maximum depths, subsample ratios, column sampling by tree, regularization alpha, regularization lambda, and minimum child samples. Ten combinations of hyperparameters were tested for each of the models.

By exploring these parameter spaces, the study aimed to identify the most effective model configurations, with the ultimate goal of maximizing AUC-ROC on the training and validation data. In case of architectures based on neural nets loss curves converged during training indicating successful learning process.

4.4. Explainability measures

This section covers the pivotal role of explainability in bridging the gap between powerful yet opaque algorithms and the trust of medical practitioners. Many advanced algorithms operate as black boxes, producing outputs that lack easily interpretable decision rules. This lack of transparency often leads to skepticism among healthcare professionals, hindering the adoption of predictive algorithms in clinical practice.

Developing techniques in explainable artificial intelligence (XAI) offer a promising solution by enabling the examination of model sensitivity to different features and providing insights into their behavior. By utilizing XAI techniques, such as those offered by SHAP (Lundberg and Lee, 2017) and Dalex (Baniecki et al., 2021) Python libraries, it becomes possible to increase trust in predictive algorithms within clinical practice. These techniques allow for a deeper understanding of how models make predictions, thereby enhancing transparency and facilitating the acceptance of AI-driven decision-making processes.

In the context of this study XAI techniques were applied to three models: top-performing CatBoost as well as LSTM and TCN.

XAI techniques can be broadly categorised into model-based explanations and local explanations (Vilone and Longo, 2021 describe it as "scope of explanation"). Model-based explanations provide an overall understanding of how a model makes predictions, which is usually achieved with some aggregations or distribution analysis. Local explanations, on the other hand, focus on explaining predictions produced for selected observations.

SHAP library was used to estimate Shapley values (Shapley et al., 1953) on a sample of 2000 random observations drawn from validation set. Because neural net architectures are based on moving windows of 20 observations the contributions of features for these models were summed along the time axis to obtain one estimate per each feature. In the context of GBT models, SHAP values (estimates of Shapley values) are calculated based on raw predictions and may exceed the range of 0-1. This approach is the default recommendation of the SHAP library documentation for tree-based models.

The core idea behind Shapley based explanations of machine learning models is to use fair allocation rules from cooperative game theory to allocate model's output - understood as pay-off - among its input features, proportionally to their contributions to the total outcome.

Shapley values help explain the contribution of each feature to the prediction of a model for a specific instance. In the context of model explanations, the Shapley values for a given prediction sum up to the difference between the expected output of the model and the local output (i.e., the prediction for the specific instance).

Shapley values are constructed to distribute the difference between the local prediction and the expected output in an additive manner:

$$\sum_{i=1}^{n} \phi_i = f(x) - E(f(X)) \tag{1}$$

The expected output, $E(f(X))$, is the expected prediction of the model over all possible instances. It serves as the baseline value if we have no information about the specific instance. The local output, $f(x)$, is the prediction of the model for a specific instance x. The Shapley value, ϕ_i, represents the contribution of feature i to the difference between the expected output and the local output.

Each Shapley value, ϕ_i, is defined as the weighted sum of marginal contributions of feature i across all possible subsets of features. This can be expressed formally as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{2}$$

where N is the set of all features, S is a subset of features not including i, and $f(S \cup \{i\}) - f(S)$ is the difference between actual model's prediction and model's prediction $f(S)$ using only features in subset S. This formula ensures that the contributions are averaged over all possible subsets in a proportional manner, leading to a fair allocation that sums up to the total difference (Shapley et al., 1953).

Apart from fairness Shapley values exhibit following properties efficiency, symmetry and null player, which are desirable for explainability reasons. Efficiency ensures that the sum of the Shapley values is equal to the total value being distributed. Symmetry implies that if two features contribute equally to all subsets, they get equal Shapley values. The null player property means that features that do not change the output when added get a Shapley value of zero. For combined models, the Shapley values of the combined model are the sum of the Shapley values of the individual models.

As a supportive means of model's explanation, Dalex Python library (Dalex) was utilised. Dalex allows for calculation of feature importance and partial dependence profiles (PDP).

Feature importance was used in the drop-out loss version, which is a measure of how much the model's performance (default: AUC-ROC) decreases when a particular feature is removed across permutations of features values. The larger the drop in performance, the more important the feature is deemed to be. This approach provides an intuitive understanding of the contribution of each feature to the model's overall accuracy and reliability, which can support explanations basing on SHAP values.

The partial dependence profile (PDP) illustrates how the expected model output changes in relation to feature values.

4.5. Savings calculation

Based on the analysis conducted by Ericson et al., 2022, the cost assessment for implementing AI in sepsis diagnosis should emphasise the timing of diagnosis and the benefits of early detection, both of which are influenced by the accuracy in distinguishing true positives from false negatives.. The economic evaluation method adopted in this study aligns with these priorities, aiming to offer a straightforward estimate of the cost savings resulting from early diagnosis and the timing effect. To achieve this, specific cost assumptions were made, which are outlined in the following paragraphs, along with the methodology used to measure the economic advantage of the AI models.

Paoli et al., 2018 published a study analyzing the relationship between the costs of sepsis hospitalization and the severity and timing of intervention. The analysis utilised the Premier Healthcare Database, which represents approximately 20% of U.S. inpatient discharges from private and academic hospitals

for adult patients. Descriptive statistics were calculated for patient demographics, characteristics, and clinical and economic outcomes during the index hospitalization for a sample from 2010 to 2016. The study estimated the average cost of severe sepsis treatment to be $19,851 when present at admission and $60,672 when not present at admission. Additionaly, as per Liang et al., 2020 the average cost of a hospital stay in the United States in 2016 was $11,700. Based on this value and the costs associated with severe sepsis treatment ($19,851 when present at admission and $60,672 when not present at admission), it is possible to provide an approximation of the relative savings incurred by using the classifier as an early diagnostic tool compared to relying solely on clinical diagnosis.

Let's consider two diagnostic approaches. The first approach relies solely on clinician diagnosis and is associated with 0 relative savings. The second approach involves using a classifier to apply preventive treatment before clinical diagnosis, based on a positive label assigned by the classifier. In the application of such a classifier for the purpose of early diagnosis, there are four possible cost scenarios:

Early Detection: The patient develops sepsis during the hospital stay, and the model classifies it before the clinician. If the patient is correctly diagnosed at admission, this results in an expected saving of approximately $40,000, which is the average difference between the cost of treating severe sepsis present at admission and not present at admission.

False Positive and Waste of Resources: The patient does not develop sepsis during hospitalization, but the model incorrectly identifies them as septic. If the patient is erroneously diagnosed at admission, this results in the application of unnecessary preventive treatment and an expected loss of $12,000, which is the average cost of a hospital stay.

False Negative: The patient had sepsis during the hospital stay, and the clinician correctly identified it, but the model failed to do so (false negative). This results in no relative savings since the clinician's diagnosis prevailed.

Coherence of Diagnoses: The classification by both the clinician and the model is the same, either correctly identifying the absence of sepsis or both missing it. This also results in no relative savings.

In both the positive scenario (resulting in a $40,000 gain) an the negative scenario (resulting in a $12,000 loss), the calculations are based on the assumption that sepsis is diagnosed (or misdiagnosed) at the time of admission. However, in practice, predictions are made based on hourly measurements, and the timing of the correct or incorrect diagnosis can vary from admission to the point of clinical diagnosis.

To account for this variability, an *hour_effect* parameter was introduced. This parameter quantifies the impact of diagnosing sepsis one hour earlier or extending treatment by one hour in the case of a false positive. The *hour_effect* assumes a linear relationship between the timing of the diagnosis and the associated costs or savings.

The above assumptions resulted in an algorithm for estimating relative savings, as detailed in Algorithm 1. Here, *predictions refer* to the model outputs, *clinical_labels* represent the clinical diagnosis at a specific hour, and *recoded_labels* indicate whether the patient developed sepsis during their entire stay. The predictions treat sepsis as an absorbing state, meaning that if two consecutive predictions are positive, the patient is classified as septic for the remainder of their treatment.

Algorithm 1 Calculate Relative Savings

Require: predictions, clinical_labels, recoded_labels, hour_effect
Ensure: relative_savings
relative_savings ← 0
for $i \leftarrow 1$ to n **do**
  if predictions[i] = 1 **and** clinical_labels[i] = 0 **and** recoded_labels[i] = 1 **then**
    relative_savings += 40000$ × hour_effect
  else if predictions[i] = 1 **and** clinical_labels[i] = 0 **and** recoded_labels[i] = 0 **then**
    relative_savings −= 12000$ × hour_effect
  end if
end for
return relative_savings

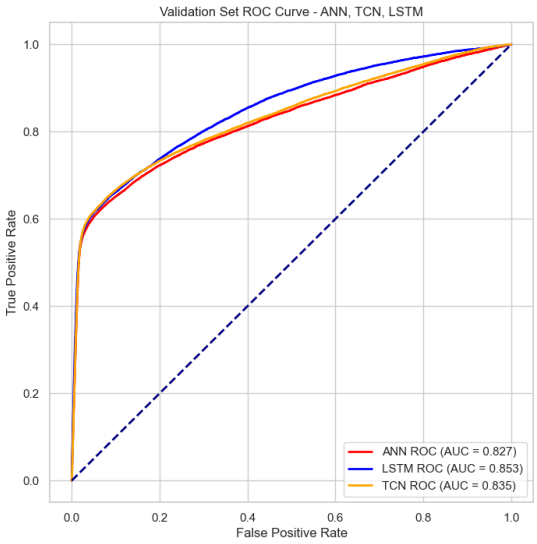
This method of calculating relative savings could be criticised for several reasons. Firstly, it only considers cases of severe sepsis, ignoring both mild sepsis and septic shock instances. This omission overlooks a significant portion of the patient population. Secondly, the standard deviation of severe sepsis costs is substantial ($75,439 for sepsis present at admission and $25,698 for sepsis not present at admission). Such high variability suggests that the mean cost may not be representative. This issue is further compounded by differences at the national level.

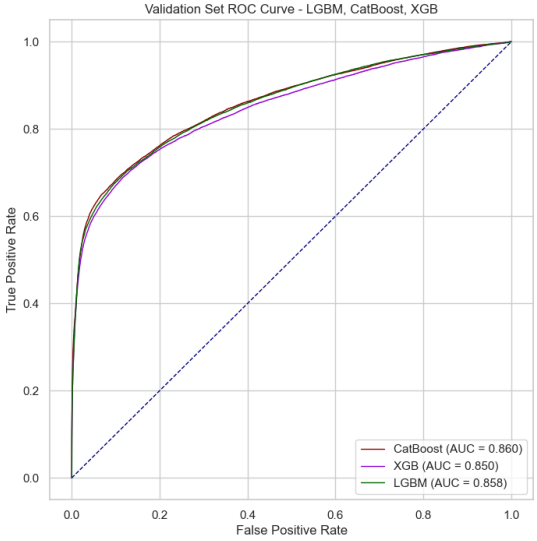
The most questionable element of proposed methodology, however, is the concept of the *hour_effect*. This proxy represents both the benefit of diagnosing sepsis one hour earlier and the cost of extending treatment for non-sepsis patients by an additional hour. Actual relationship between time and cost is likely different for sepsis-positive and sepsis-negative patients and is probably not linear, with marginal costs increasing with each additional hour of delay. Despite these limitations, for the purpose of a quick analysis providing comparative insights into different models, this rough estimate may still offer satisfactory results.

5. Results

5.1. Predictive Power

Table 2 presents the binary cross-entropy loss (reweighted for positive labels) and AUC-ROC scores for each examined model, evaluated on both the training and validation sets. The training set results are further divided into the results of training neural network models and results from cross-validation with GBT models. The corresponding ROC curves are illustrated in Figure 3. Additionally, Figure 4 displays the ROC curves for the test data, showcasing the performance of the top-performing models on an independent sample that was neither used for training nor for optimization.





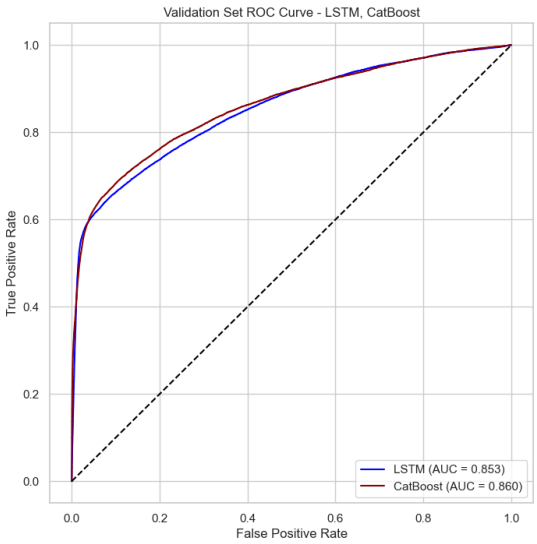


Figure 3: Area under ROC for best models in each class on validation set.
Source: Own preparation.

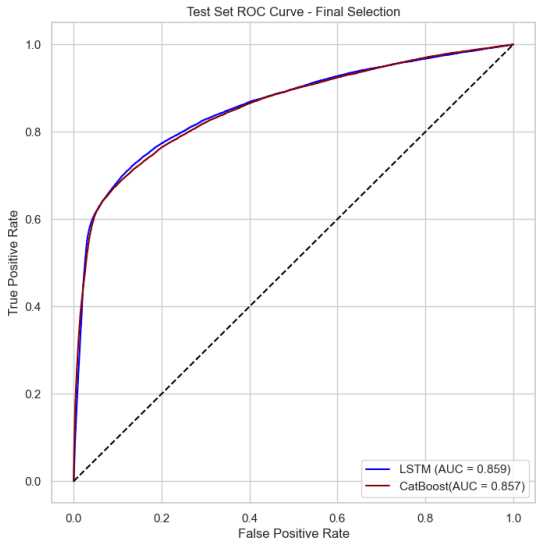


Figure 4: Area under ROC for best LSTM and CatBoost on test set.
Source: Own preparation.

As can be seen from the Table 2, the most performing neural network based architecture was the LSTM, achieving a training AUC-ROC of 0.863 and a validation AUC-ROC of 0.853. It is important to note that the values of the loss associated with neural nets training cannot be interpreted in absolute terms due to *pos_weight* parameter, which was introduced to tackle class imbalance.

Nevertheless, it can be observed that for ANN, TCN, and LSTM, the loss reached comparable levels between validation and training data indicating their stability. The difference between validation and training loss was highest for LSTM (0.06).

Model	**Loss**		**AUC-ROC**		
	Train	**Val**	**Train**	**Train (CV)**	**Val**
ANN	0.99	0.96	0.831	-	0.827
TCN	0.96	0.99	0.834	-	0.835
LSTM	0.91	0.97	0.863	-	0.853
CatBoost	-	-	-	0.786	0.860
XGBoost	-	-	-	0.805	0.850
LGBM	-	-	-	0.787	0.858

Table 2: Training results.
Source: Own preparation.

GBT models were also evaluated with respect to training and validation sets. The AUC-ROC score for the training set was computed as the average score achieved across each of the five folds during cross-validation (CV). Although it could be argued that XGBoost performed best during CV, CatBoost, with its highest score on the validation set as selected as the best architecture among tree-based architectures.

Figure 4, which depicts the ROC curve for the final selected models (CatBoost and LSTM) on test set, further validates their performance. Notably, both models exhibit stable results across the validation and test sets. The LSTM model performed better on the test set (0.859) compared to the validation set, while

the CatBoost model showed a slight decline in performance on the test set (0.857). Consequently, the slight difference between the models' AUC-ROC decreased from 0.007 to 0.002 favouring LSTM on the test set. Nevertheless, the final economic assessment regarding the attained savings level, along with the explainability measures and time efficiency of training, indicate that CatBoost was the most adoptable among considered models.

The results displayed in the figures could be compared to those demonstrated within the literature.

(Mao et al., 2018) used only six vital signs (heart rate, respiratory rate, oxygen saturation, temperature, systolic, and diastolic blood pressure) to train a GBM model, achieving an AUC-ROC of 0.83. The three most performing models from the PhysioNet Challenge, from which training data was derived, scored AUCs of 0.868, 0.863, and 0.833, respectively. When comparing these numbers with score obtained by models presented in this study, it is important to note that the objective of the challenge differed from this study, as classifiers predicting sepsis twelve hours before the actual onset were penalised.

In an independent project (Bobra, 2021) implementing an ensemble that combined LSTM with an ANN network with the same objective and with the use of the data from PhysioNet Challenge, the AUC-ROC was 0.77 and 0.76 for validation and test sets, respectively.

The crucial reference which served as an inspiration for this study is the XAI-EWS scoring system, trained on restricted data from a Danish cohort of patients. This paper reported the AUC-ROC of the model for sepsis to be between 0.8 and 0.92 for each of the five folds of cross-validation - value comparable to those achieved by both GBT and neural net models.

The second main goal of this paper was to evaluate the performance of neural nets, with a specific focus on TCN. The results reveal that LSTM networks, a traditional approach to modelling sequences, outperformed TCN. Additionally, GBT architectures generally yielded better results than neural networks, with CatBoost emerging as the top-performing model on validation set and in economic evaluation (Section 5.3).

5.2. Explainability

In the subsequent section, performance of the models is evaluated with respect to the second critical aspect of ML algorithm's adoptability: explainability.

Beeswarm plots (Figures 5, 6 and 7 for each of the models display the distributions of SHAP values with respect to feature levels. Level 0 indicates expected value of model output calculated on the validation sample: 0.38 for LSTM, 0.48 for TCN and 0.456 for CatBoost. Examining beeswarm plots reveals clear similarities between the contributions of features among the models.

First of all ICULOS signifying ICU length of stay is the most important and dominating feature for both TCN, LSTM and CatBoost. ICULOS is a critical feature because it directly relates to the severity of a patient's condition. Longer ICU stays typically indicate more severe illnesses or complications, which naturally affects predictions related to patient outcomes.

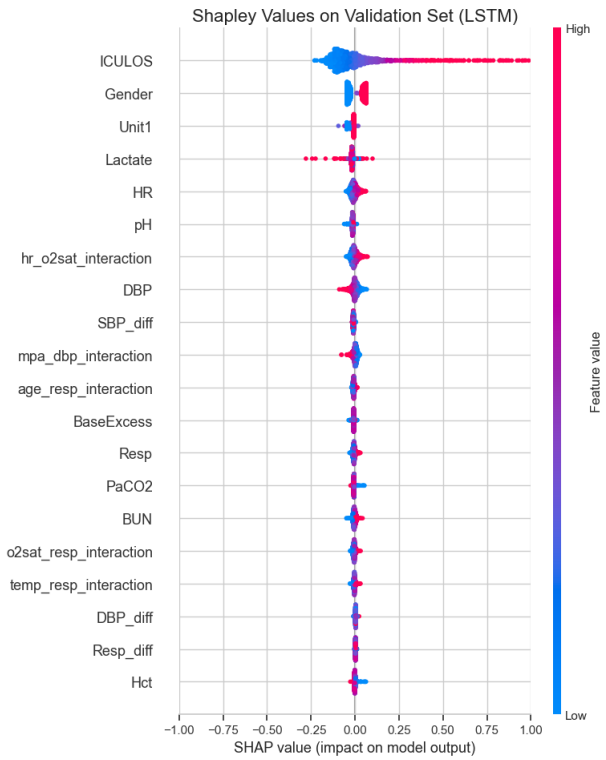


Figure 5: Shapley Values for LSTM on Test Set.
Source: Own preparation.

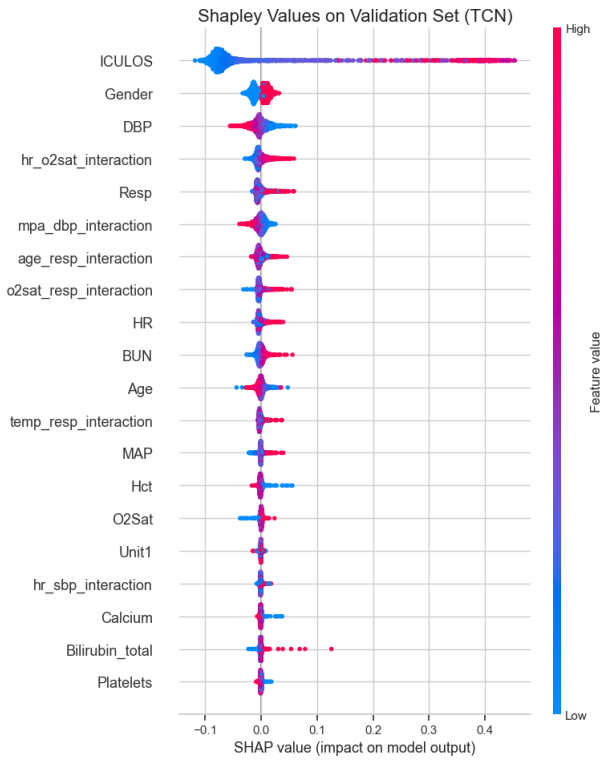


Figure 6: Shapley Values for TCN on Test Set.
Source: Own preparation.

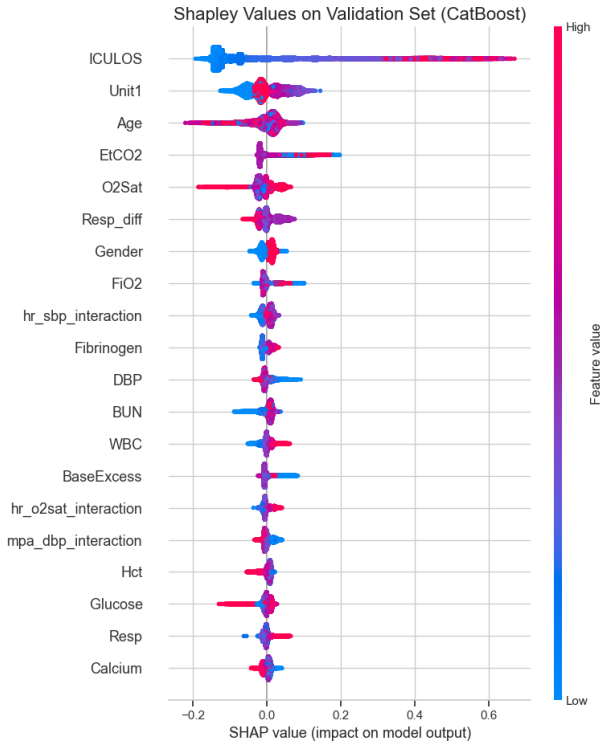


Figure 7: Shapley Values for CatBoost on Validation Set.
Source: Own preparation.

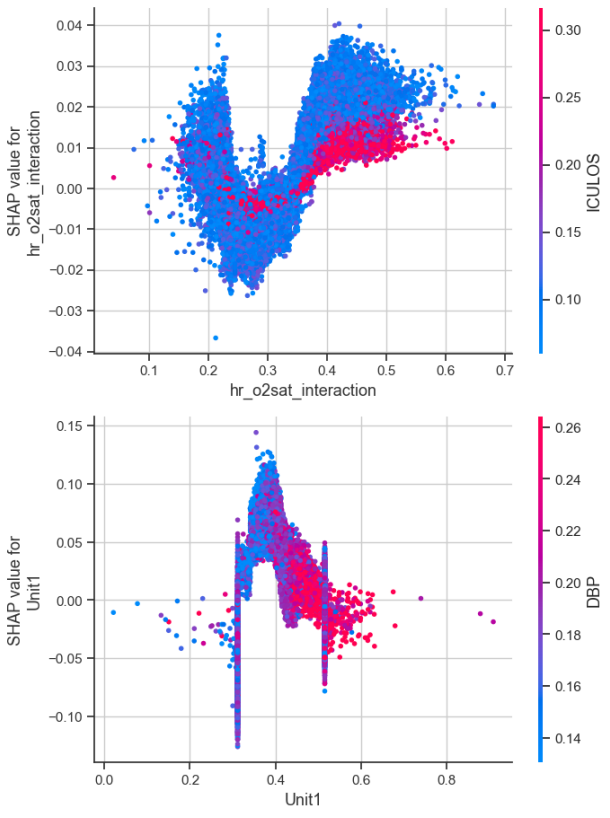


Figure 8: Dependency Plots for CatBoost.
Source: Own preparation.

ICU length of stay is also considered an important metric in understanding septic shock Tuttle et al., 2023 and patients mortality (Rodrigues et al., 2024).

The impact of ICULOS is also illustrated in the dependency plot for the top-performing model, CatBoost, as shown in Figure 8. This plot allows us to study the relationship between SHAP values, the interaction of heart rate and oxygen saturation, and ICULOS. The positive correlation confirms the importance of heart rate and oxygen saturation as indicators for sepsis detection. Additionally, the cluster of red points, slightly elevated, indicates that a longer length of stay is usually associated with a higher chance of sepsis, as well as elevated levels of oxygen saturation or heart rate.

Gender was the second most contributing variable for LSTM and TCN. Gender differences profoundly impact disease prevalence, progression, treatment response, and recovery. As noted by Lakbar et al., 2023, sepsis and septic shock are more prevalent in men than in women. This disparity is evident in the distribution of SHAP values: for all three models, the variable's effect shows a positive contribution for being male and negative for being female.

The localisation factor expressed by Unit1 variable was less significant for TCN. It could mean that TCN was more sensitive to biomarkers, or that it was not able to capture the effect of hospital conditions - conversely to LSTM and CatBoost. The latter were more sensitive to this variable as it was the third most significant variable for them as per beeswarm plots.

The general dependency plot for CatBoost (Figure 8) illustrates the effect of the Unit1 variable, which was imputed for 39% of the data. For patients in the first hospital, where Unit1 values are around 0.3, the distribution is more dispersed with a peak around 0.4. In contrast, for patients in the second hospital, where Unit1 values are closer to 0.5, the values are more concentrated.

When focusing on the interval from 0.3 to 0.5, a gradual decrease in SHAP values can be observed as the probability of being a patient at the second hospital increases. This decrease in sepsis risk is accompanied by an increase in Diastolic Blood Pressure (DBP), which aligns with the fifth hypothesis.

It could be generally noted that variables with higher missingness level (apart from biliurbin and hematrocit) were less informative for the models and thus they were not included in the top 20 significant features. In case of variables which proved informative the models outcomes were in line with expectations, which could be formed basing on medical knowledge. The directions of relationships between the models' outcomes and feature values were also aligned, affirming the validity of the models. One noteworthy difference that required further investigation is the significance of EtCO2 in CatBoost, which distinguishes it from neural network-based models. More detailed explanations regarding EtCO2, and contributions of other significant variables, are provided in the following sections.

Variables with positive influence:

Apart from the three navigating features one group of variables had a strict positive influence on the models' outcome.

Amid these there was: heart rate, interaction of heart rate and oxygen saturation, respiration, interaction of age and respiration (in case of TCN), oxygen saturation, bilirubin (in case of TCN), BUN and mean arterial pressure. The positive impact of Fibrinogen and WBC (white blood count) was observable and strong in case of CatBoost.

Heart rate is a significant predictor because an elevated heart rate, or tachycardia, is a common early sign of sepsis. When the body fights an infection, it often increases the heart rate to maintain adequate perfusion despite the systemic inflammation caused by the infection. The importance of HR as a marker of sepsis onset and progression is also acknowledged in literature (Shashikumar et al., 2017).

The interaction of heart rate and oxygen saturation is also important. This interaction reflects the body's oxygen delivery and utilization efficiency. In sepsis, the body's demand for oxygen increases while its ability to transport and utilise oxygen may be compromised due to systemic inflammation and potential organ dysfunction. This can lead to further complications, making this interaction a vital predictor of sepsis.

Respiratory rate is another critical feature. In sepsis, the respiratory rate often increases (Lee et al., 2021) as the body tries to meet the higher metabolic demands and compensate for potential respiratory complications, such as acute respiratory distress syndrome (ARDS). Rapid breathing helps to increase oxygen intake and expel carbon dioxide, but it can also indicate the severity of the sepsis, thus increasing its predictive value.

The interaction of age and respiration is particularly relevant in the case of TCN models. Age influences the physiological response to sepsis, as older patients often have a different respiratory response compared to younger individuals. Within period analysed in study conducted by Martin et al., 2006, elderly patients (over 65 years of age) accounted for 12% of the U.S. population, but 64.9% of sepsis cases.

Oxygen saturation, typically regarded as an indicator of good health, can also reflect increased tissue demand and impaired oxygen extraction. According to Textoris et al., 2011, excessively high oxygen saturation is associated with a decreased survival rate and increased severity of sepsis. Although this observation contrasts with some views in the literature, it is further confirmed by the consistent positive effect of interactions between oxygen saturation and both heart rate and respiratory rate.

Bilirubin levels, significant in the context of TCN models, can indicate liver dysfunction. Elevated bilirubin is a sign of liver stress or failure, which is a common complication in sepsis due to systemic inflammation and poor perfusion. Bilirubin is a reported marker of sepsis severity and associated mortality (Patel et al., 2015).

Mean arterial pressure is essential for maintaining adequate blood flow and organ perfusion. In sepsis, blood pressure can drop significantly due to systemic vasodilation, leading to septic shock if not managed properly. However high mean arterial pressure can mark organism effort to maintain the level of perfusion.

Elevated BUN (Blood Urea Nitrogen) levels can indicate impaired kidney function, which is common in septic patients due to low perfusion and systemic inflammation. High BUN levels reflect the kidneys' decreased ability to filter urea from the blood, and are used in predicting sepsis-associated mortality (Wang et al., 2023).

In sepsis, white blood count (WBC) levels can be either elevated or decreased, reflecting the body's reaction to infection. An elevated WBC typically indicates an ongoing infection and can be used as a predictor of sepsis-associated mortality (Rimmer et al., 2022). Conversely a decreased WBC might suggest severe sepsis or septic shock.

Fibrinogen is a vital protein involved in blood clotting. In septic patients, fibrinogen levels can be significantly affected. Elevated fibrinogen levels can indicate an acute phase response to infection, where the body is attempting to manage widespread inflammation (Tsantes et al., 2023).

The third hypothesis posited that certain variables would increase the risk of sepsis, specifically: ICULOS, age, male gender, respiratory rate, heart rate, increased leukocyte levels (expressed by the WBC variable), as well as indicators of kidney or liver dysfunction (BUN, bilirubin). SHAP values analysis revealed a consistent positive effect of these variables on the model outputs. These effects were consistent across the TCN, LSTM, and CatBoost models.

Variables with negative influence:

Another group of variables had a clear negative influence on the models' outcome. Among these: diastolic blood pressure, mean arterial and diastolic blood pressure interaction, hematrocit (in case of TCN and CatBoost), calcium (in case of TCN), platelets (in case of TCN), PaCo2 (in case of LSTM and CatBoost) and FiO2 (in case of CatBoost).

Diastolic blood pressure (DBP) is a critical feature because higher values often indicate better overall cardiovascular health and adequate perfusion. In the context of sepsis, maintaining higher diastolic blood pressure can be a protective factor against shock and organ failure, as it suggests the body's ability to maintain sufficient blood flow, whearas lower DBP can be indicative of sepsis and increase mortality (Gao et al., 2023). The normal level of diastolic blood pressure is below 80. The interaction between mean arterial pressure and diastolic blood pressure further reflects this effect. When both measures indicate robust circulatory function, it implies that the patient's cardiovascular system is effectively managing blood flow and pressure.

Hematocrit (HCT) levels, representing the proportion of red blood cells in the blood. In the context of sepsis, a higher HCT can indicate better oxygenation and a lower risk of hypoxia and related complications, thus decreasing the probability of sepsis, while reduction of HCT is characteristic of sepsis (Agnello et al., 2021).

Calcium levels, significant for TCN, are crucial for numerous cellular functions. Adequate calcium levels can help maintain cellular integrity and function, which is vital for mounting an effective response to infection. Conversely, low ionised calcium level is an independent predictor of mortality to severe sepsis (Cekmen et al., 2021).

Platelet count is another significant variable for predicting

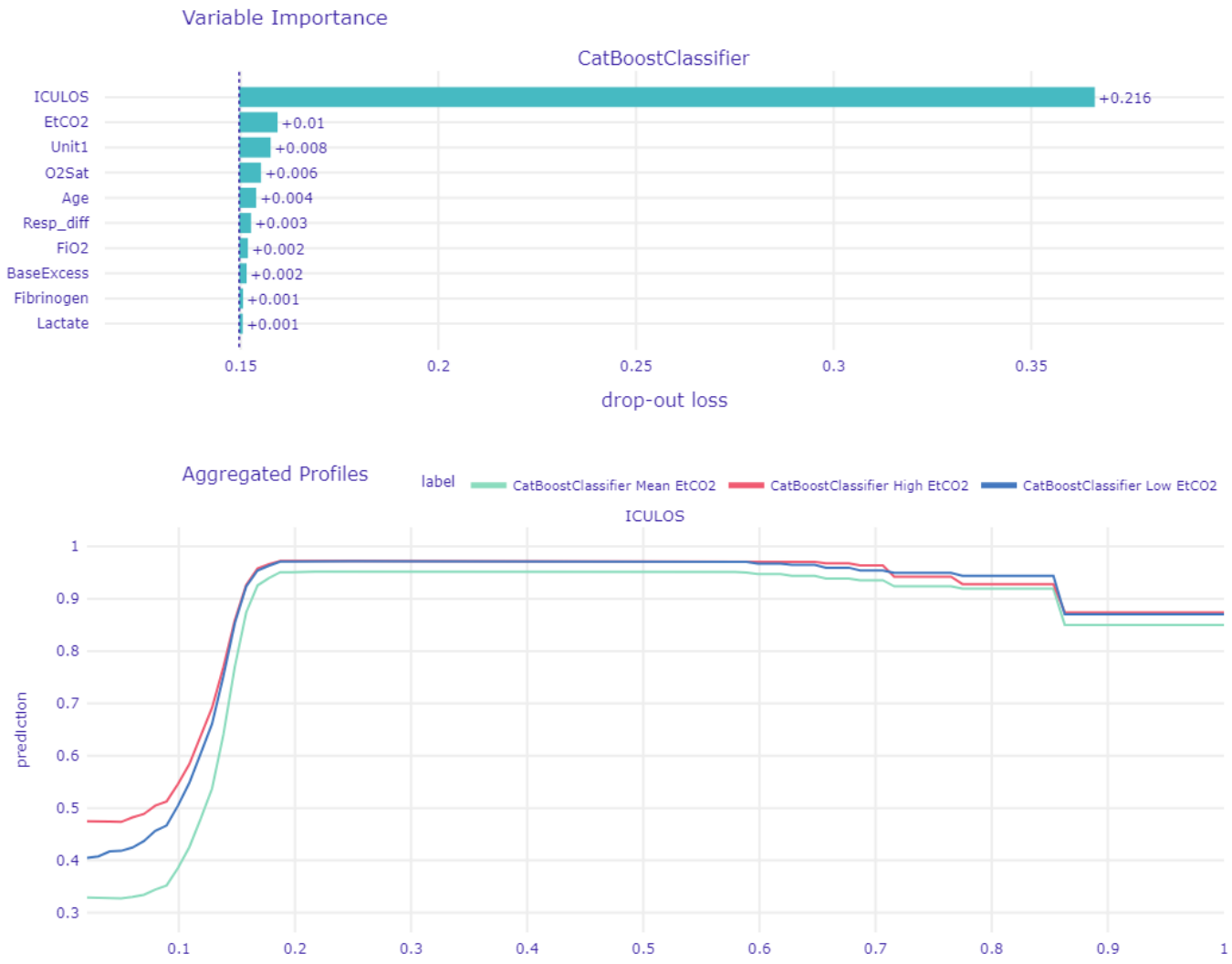


Figure 9: Feature importance and Partial Dependence Profile for CatBoost.
Source: Own preparation.

sepsis and sepsis mortality (Hua et al., 2023). Higher platelet counts can indicate a better state of health, as platelets play a vital role in hemostasis and the immune response.

The partial pressure of carbon dioxide (PaCO2),which was significant for LSTM, is a measure of respiratory function. Normal PaCO2 levels indicate effective ventilation and respiratory efficiency, which are crucial for maintaining acid-base balance and oxygenation. Both low and high levels of this biomarker can be indicative of sepsis-associated mortality (Qu et al., 2023).

FiO2, or the fraction of inspired oxygen, is a measure of the oxygen concentration being delivered to a patient. Decreased FiO2 levels are often symptomatic for patients experiencing respiratory distress. PaCO2 divided by FiO2 is a well-known parameter to assess respiratory dysfunction, used in Sequential Organ Failure Assessment and predicting sepsis mortality (Santana et al., 2013, Bi et al., 2023).

The fourth hypothesis expected variables indicative of good health and physiological efficiency (hemoglobin/hematocrit levels, diastolic blood pressure, platelet count) to decrease the risk of sepsis. This expectation was confirmed by the results from all models. Thus, it can be concluded that all models aligned with medical intuitions regarding both positive and negative influence of variables on model output.

CatBoost-specific explanations:

With respect to best performing model - CatBoost - Dalex (Baniecki et al., 2021) library was utilised as another means of model's explanation. Dalex supports calculation of feature importance and partial dependence profiles (PDP).

Top 10 feature importance calculated for variables are demonstrated on Figure 9.

Once again it can be noticed that ICULOS was the most important variable in prediction with over 0.21 drop in AUC-ROC

after removing it from dataset. The localisation effect expressed by Unit1 variable was also taken into consideration. The variables which were disregarded by both neural net architectures, but were relatively important for CatBoost predictions were: the difference in respiration rate, and most prominently end-tidal carbon dioxide (EtCO2). EtCO2 importance distinguished CatBoost from neural net based models also during SHAP values examination.

Resp_diff, the difference in respiration rates between timestamps, can reflect changes in a patient's respiratory status. Rapid changes in respiration rate might indicate worsening respiratory function or distress, which can be associated with sepsis.

ETCO2, or end-tidal carbon dioxide, measures the concentration of CO2 at the end of an exhaled breath, reflecting ventilation efficiency and, indirectly, metabolic activity. Low ETCO2 can indicate poor perfusion and metabolic acidosis, common in severe sepsis and septic shock.

The partial dependence profile (PDP) was presented as the second plot on Figure 9. ICULOS was selected as the analysed feature, with additional grouping by EtCO2.

EtCO2 was categorised into three groups based on its deviation from the mean: values below one standard deviation from the mean were represented as $mean(EtCO2) - 3 * std(EtCO2)$, values above one standard deviation from the mean were represented as $mean(EtCO2) + 3 * std(EtCO2)$, and all others were represented as $mean(EtCO2)$. This categorization facilitates an analysis of how EtCO2 affects the model's profile for ICULOS.

The relationship between model outputs and ICU length of stay reveals a clear pattern. Patients who left the ICU within a few hours exhibited a low risk of sepsis. However, the risk of sepsis increased significantly with longer stays, particularly those extending beyond 48 to 72 hours, where the risk remained consistently high. Interestingly, for the longest stays, the risk began to decrease again.

The influence of end-tidal carbon dioxide (EtCO2) on the predicted sepsis risk exhibits a non-linear relationship. Patients with elevated EtCO2 levels show the highest risk of sepsis. Conversely, those with EtCO2 levels near the mean demonstrate the lowest risk, even in the later stages of hospitalization. Patients with low EtCO2 values also demonstrate high risk of sepis, yet lower than for increased EtCO2. In the later phases the expected model output for low and high levels of EtCO2 is indistinguishable.

Long and Dippenaar, 2022 conducted a summary of literature on the influence of EtCO2 on sepsis probability and severity. Studies have found that EtCO2 levels decline in the setting of poor perfusion, indicating that low EtCO2 values may elevate the probability of sepsis. Additionally, there is a recognised strong linear relationship between EtCO2 and HCO3 levels. A decrease in HCO3 concentration is indicative of metabolic acidosis, which is commonly associated with sepsis.

On the other hand, some studies Bindu et al., 2020 and Tautz et al., 2010 suggest that increased levels of EtCO2, indicative of heightened metabolism and fever, may also be associated with sepsis. Therefore, the model's two-way, non-linear response to EtCO2 levels appears appropriate.

5.3. Economic Advantage

This section further evaluates the adoptability of the top-performing models by analyzing the costs and benefits of their potential application in a practical setting.

The savings assessment algorithm discussed in (4.5) was initially applied to optimise the thresholds of the two models identified as the most performing in their class (LSTM and CatBoost) on the validation set. The optimisiation was performed with respect to obtained levels of savings.

The assumed value of the *hour_effect* for threshold optimization was 0.02. Program iterated through various threshold values in increments of 0.01, calculating savings at each step. As demonstrated in Table 3, the CatBoost model outperformed the LSTM model in terms of economic advantage by significant margin.

Model	**Best Threshold**	**Savings at Threshold**	
		Best	**0.5**
LSTM	0.74	6,709,920	3,350,720
CatBoost	0.83	6,972,000	4,006,240

Table 3: Results of threshold optimisation on validation set ($).
Source: Own preparation.

It is important to note that optimizing the threshold resulted in a substantial increase in savings, doubling the savings for the LSTM model and nearly doubling them for the CatBoost model. This optimization involved setting the threshold above the default level of 0.5, thereby making the models more selective for the positive class.

This outcome, driven by economic optimization, may not align with the primary medical objective of prioritizing patient life and health. However, it is important to remember that this algorithm serves as a supportive diagnostic tool rather than a definitive solution. The economic scenario for cost calculation is based on the assumption that clinicians' assessments are the decisive factor, with the models providing most prudent recommendation for earlier intervention. In a scenario maximising performance of the model the threshold guaranteeing optimal balanced accuracy of other metric of choice could be considered.

Hour Effect	**CatBoost**	**LSTM**
Relative Savings on Validation Set		
0.005	1,743,000	1,677,480
0.01	3,486,000	3,354,960
0.02	6,972,000	6,709,920
0.05	17,430,000	16,774,800
Relative Savings on Test Set		
0.005	1,789,720	1,756,640
0.01	3,579,440	3,513,280
0.02	7,158,880	7,026,560
0.05	17,897,200	1,7566,400

Table 4: Cost and relative savings ($).
Source: Own preparation.

After determining the optimal thresholds, it became possible to conduct a simulation assessing the impact of varying *hour_effect* values on relative savings for both models across validation and test sets. The results of this simulation are presented in Table 4. Both the validation and test sets contained an equal number of patients, each comprising **6,051** individuals. The *hour_effect* values tested ranged from 0.005 to 0.05.

As can be seen from above CatBoost model demonstrated the greatest increase in relative savings, outperforming LSTM on both validation and test sets.

6. Summary and Conclusions

The work described in this article investigates the potential of ensuring three key characteristics that facilitate the adoption of an AI-DDS system: predictive power, explainability, and economic advantage. The article revisits achievements already documented in the literature and then, after discussing the methods, moves to sections devoted to evaluating each of these characteristics on proposed practical example.

The methods section outlined the details of model development and evaluation. These included the model selection process, the quality and representativeness of the dataset used for training and testing, and the architectures applied. It also covered the measures of explainability and the algorithm used to assess economic advantage. The discussion addressed these measures from both a theoretical perspective, providing necessary context, and a practical standpoint, detailing the specific libraries and custom solutions.

After discussing methods, the results of model training and assessment were presented. Firstly the evaluation of the discriminative power of models with AUC-ROC on training, validation, and test sets was described, followed by examination of selected models' explainability.

XAI analysis allowed to verify that directions of relationships between the model outcomes and feature values were aligned with expectations based on medical knowledge. It could also be observed that variables with higher levels of missingness (apart from bilirubin levels and hematocrit) were less informative for the models and thus were not included in the top 20 significant features displayed on plots.

Lastly, potential savings from utilisation of trained algorithms in sepsis diagnosis were evaluated.

By these means, the first aim of research was realised with two models adhering to requirements for predictive power, explainability and economic advantage. Among tested models, CatBoost and LSTM achieved the best discriminative power, however, CatBoost proved superior in terms of estimated economic savings.

The secondary aim of research was to validate the application of Temporal Convolutional Network (TCN), proposed as the most performing algorithm in the Explainable AI Early Warning Score (XAI-EWS) system (Lauritsen et al., 2020). Validation was performed on a Physionet dataset by comparing TCN against LSTM and best GBT model to determine if the TCN-based XAI-EWS system improves predictive performance across diverse clinical data. TCN was outperformed by both LSTM and CatBoost.

The third objective was to validate the trained models against established medical knowledge. Expectations were defined for two groups of variables, predicting either positive or negative impacts on sepsis risk. With the help of explainability techniques, it was possible to show that these expectations were consistently met across all models without any violations. This alignment with medical knowledge strengthens the credibility and reliability of the models' predictions.

The fourth aim of this article was to propose a satisfactory approach for assessing the economic value of the most performing models. Thanks to several assumptions simplifying cost estimation and allowing for comparison, the financial effect of early treatment due to algorithmic support was calculated. CatBoost achieved the highest level of savings outperforming the second-best model, LSTM.

Research goals and expectations regarding the medical interpretability of the model outcomes based on literature led to the formulation of five research hypotheses. With respect to these, following conclusions can be drawn:

Available open-source algorithms, with appropriate preparation, can indeed meet the requirements for AI-DDS. Thus, the first hypothesis was confirmed.

Temporal Convolutional Network (as proposed by the authors of the XAI-EWS system) did not outperform all tested models, falling short of LSTM and CatBoost variants. Thus, the second hypothesis was verified negatively.

ICULOS, age, male gender, and variables indicating increased metabolism or ongoing immune response of the organism (respiratory rate, heart rate, increased leukocyte levels) had a significant positive impact on models' output. Thus, the third hypothesis was confirmed.

Variables defining good condition and efficiency of the organism (hemoglobin/hematocrit, diastolic blood pressure, platelets count) had a significant negative impact on models' output. Therefore, fourth hypothesis was confirmed.

The application of all algorithms to diagnose patients from the test sample would lead to significant cost savings. Therefore, the fifth hypothesis was also confirmed.

It is important to note that the third and fourth hypotheses could only be verified through the use of XAI techniques. This highlights the crucial role of explainability in bridging the gap between the power of black-box algorithms and the trust of medical practitioners, offering valuable insights into the algorithm's inner workings. For instance, a detailed explanation revealed the non-linear influence of EtCO2 on sepsis risk: both abnormally low and high EtCO2 levels were associated with an increased risk of sepsis, while levels closer to the mean were linked to a reduced risk.

XAI techniques, such as those offered by SHAP and Dalex Python libraries, not only allow for a deeper understanding of how models make predictions. Moreover, the adoption of XAI can help sharpen and unify definitions of illnesses by providing clearer insights into the underlying factors influencing disease outcomes. An example of such insight can be highlighted also

with respect to this study: consistent interpretation of oxygen saturation as a factor increasing sepsis risk - across independently trained models - addresses ongoing debates in the literature about its role in sepsis.

One limitation of this study is the dataset used, which contains a significant amount of missing data. In real-world settings, access to more comprehensive and higher-quality datasets could create a more favorable environment for model development and potentially enhance the predictive power of the AI models. This limitation underscores the need to validate the proposed solutions against a broader range of data to ensure their robustness and generalizability.

Most recent challenges in AI adoption for sepsis diagnosis include the need for local-level, actionable explanations and early sepsis detection (Kamran et al., 2024, Zhang et al., 2024). While this study puts emphasis on early sepsis prediction throughout patient's entire hospital stay, with respect to XAI it primarily focuses on model-specific explanations and does not sufficiently tackle the issue of local-level explanations. Limited focus on patient-specific predictions may restrict the practical applicability of AI models in clinical decision-making.

Finally, presented assessment of economic advantage is somewhat constrained. The simplifications made in cost estimation might overlook the complex interactions between various factors influencing clinical decisions. A more thorough practical evaluation using a controlled sample of patients could provide a more accurate understanding.

In addressing the limitations related to dataset quality, future studies should aim to validate the models using more diverse and comprehensive datasets. This could involve collaborations with healthcare institutions to obtain real-world data, ensuring that the models are tested in varied and realistic scenarios.

Future research should also focus on enhancing the XAI assessment system to account for local predictions, providing more granular insights that are essential for medical diagnostics. This development would involve extending the proposed XAI techniques to offer more detailed explanations at the individual patient level, thereby improving the clinical utility of the models.

Additionally, the economic evaluation could be enhanced by incorporating more complex interactions between factors influencing clinical decisions, or by developing a dedicated model specifically for cost assessment. Conducting real-world simulations would certainly improve the precision of cost savings estimates related to early sepsis diagnosis, offering more robust evidence for the economic viability of AI-driven decision support systems in healthcare.

References

Agnello, L., Giglio, R.V., Bivona, G., Scazzone, C., Gambino, C.M., Iacona, A., Ciaccio, A.M., Lo Sasso, B., Ciaccio, M., 2021. The value of a complete blood count (cbc) for sepsis diagnosis and prognosis. Diagnostics 11, 1881.

Avendaño-Ortiz, J., Maroun-Eid, C., Martín-Quirós, A., Lozano-Rodríguez, R., Llanos-González, E., Toledano, V., Gómez-Campelo, P., Montalbán-Hernández, K., Carballo-Cardona, C., Aguirre, L.A., et al., 2018. Oxygen saturation on admission is a predictive biomarker for pd-l1 expression on circulating monocytes and impaired immune response in patients with sepsis. Frontiers in Immunology 9, 2008.

Aygun, U., Yagin, F.H., Yagin, B., Yasar, S., Colak, C., Ozkan, A.S., Ardigò, L.P., 2024. Assessment of sepsis risk at admission to the emergency department: Clinical interpretable prediction model. Diagnostics 14, 457.

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P., 2021. dalex: Responsible machine learning with interactive explainability and fairness in python. Journal of Machine Learning Research 22, 1–7. URL: `http://jmlr.org/papers/v22/20-1473.html`.

Bi, H., Liu, X., Chen, C., Chen, L., Liu, X., Zhong, J., Tang, Y., 2023. The pao2/fio2 is independently associated with 28-day mortality in patients with sepsis: a retrospective analysis from mimic-iv database. BMC Pulmonary Medicine 23, 187.

Bindu, B., Singh, G.P., Jain, V., Chaturvedi, A., 2020. A persistently high end-tidal carbon dioxide value: can this be spurious? Journal of Neuroanaesthesiology and Critical Care 7, 104–106.

Bobra, N., 2021. `https://github.com/nerajbobra/sepsis-prediction/tree/master`.

Cardoso, L.T., Grion, C.M., Matsuo, T., Anami, E.H., Kauss, I.A., Seko, L., Bonametti, A.M., 2011. Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. Critical care 15, 1–8.

Cekmen, B., Koylu, R., Akilli, N.B., Gunaydin, Y.K., Koylu, O., Atis, S.E., Cander, B., 2021. Ionized calcium level predicts in-hospital mortality of severe sepsis patients: A retrospective cross-sectional study. Journal of Acute Disease 10, 247–251.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

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, ACM, New York, NY, USA. pp. 785–794. URL: `http://doi.acm.org/10.1145/2939672.2939785`, doi:`10.1145/2939672.2939785`.

Dalex, . `https://pypi.org/project/dalex/`.

Dorogush, A.V., Ershov, V., Gulin, A., 2018. Catboost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 .

Ericson, O., Hjelmgren, J., Sjövall, F., Söderberg, J., Persson, I., 2022. The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in sweden. Journal of health economics and outcomes research 9, 101.

Escobar, G.J., Liu, V.X., Schuler, A., Lawson, B., Greene, J.D., Kipnis, P., 2020. Automated identification of adults at risk for in-hospital clinical deterioration. New England Journal of Medicine 383, 1951–1960.

Fan, W., Liu, J., Zhu, S., Pardalos, P.M., 2020. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (aimdss). Annals of Operations Research 294, 567–592.

Fathi, M., Markazi-Moghaddam, N., Ramezankhani, A., 2019. A systematic review on risk factors associated with sepsis in patients admitted to intensive care units. Australian Critical Care 32, 155–164.

Gao, Z., Li, C., Chen, H., Chen, D., Ma, S., Xie, J., Wu, C., Liu, L., Yang, Y., 2023. Association between diastolic blood pressure during the first 24 h and 28-day mortality in patients with septic shock: a retrospective observational study. European journal of medical research 28, 329.

González-Vidal, A., Rathore, P., Rao, A.S., Mendoza-Bernal, J., Palaniswami, M., Skarmeta-Gómez, A.F., 2020. Missing data imputation with bayesian maximum entropy for internet of things applications. IEEE Internet of Things Journal 8, 16108–16120.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hua, Y., Wang, R., Yang, J., Ou, X., 2023. Platelet count predicts mortality in patients with sepsis: A retrospective observational study. Medicine 102, e35335.

IQBAL, A., Sikdar, B., 2023. Are classifiers trained on synthetic data reliable? an xai study. Authorea Preprints .

Kamran, F., Tjandra, D., Heiler, A., Virzi, J., Singh, K., King, J.E., Valley, T.S., Wiens, J., 2024. Evaluation of sepsis prediction models before onset of treatment. NEJM AI 1, AIoa2300032.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30, 3146–3154.

Khanna, N.N., Maindarkar, M.A., Viswanathan, V., Fernandes, J.F.E., Paul, S., Bhagawati, M., Ahluwalia, P., Ruzsa, Z., Sharma, A., Kolluri, R., et al.,

2022. Economics of artificial intelligence in healthcare: diagnosis vs. treatment, in: Healthcare, MDPI. p. 2493.

Kim, H.I., Park, S., 2019. Sepsis: early recognition and optimized treatment. Tuberculosis and respiratory diseases 82, 6–14.

Lakbar, I., Einav, S., Lalevée, N., Martin-Loeches, I., Pastene, B., Leone, M., 2023. Interactions between gender and sepsis—implications for the future. Microorganisms 11, 746.

Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B., 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications 11, 3852.

Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165.

Lee, C.U., Jo, Y.H., Lee, J.H., Kim, J., Park, S.M., Hwang, J.E., Lee, D.K., Park, I., Jang, D.H., Lee, S.M., 2021. The index of oxygenation to respiratory rate as a prognostic factor for mortality in sepsis. The American journal of emergency medicine 45, 426–432.

Li, M., Huang, P., Xu, W., Zhou, Z., Xie, Y., Chen, C., Jiang, Y., Cui, G., Zhao, Q., Wang, R., 2022. Risk factors and a prediction model for sepsis: A multicenter retrospective study in china. Journal of Intensive Medicine 2, 183–188.

Liang, L., Moore, B., Soni, A., 2020. National inpatient hospital costs: The most expensive conditions by payer, 2017. Rockville (MD): Agency for Healthcare Research and Quality (US); 2006 Feb-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK561141/.

Long, J., Dippenaar, E., 2022. To what extent is end-tidal carbon dioxide a predictor of sepsis? Journal of Paramedic Practice 14, 425–431.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

M. Mostafa, S., S. Eladimy, A., Hamad, S., Amano, H., 2020. Cbrl and cbrc: Novel algorithms for improving missing value imputation accuracy based on bayesian ridge regression. Symmetry 12, 1594.

Mao, Q., Jay, M., Hoffman, J.L., Calvert, J., Barton, C., Shimabukuro, D., Shieh, L., Chettipally, U., Fletcher, G., Kerem, Y., et al., 2018. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. BMJ open 8, e017833.

Martin, G.S., Mannino, D.M., Moss, M., 2006. The effect of age on the development and outcome of adult sepsis. Critical care medicine 34, 15–21.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5, 115–133.

Morrill, J., Kormilitzin, A., Nevado-Holgado, A., Swaminathan, S., Howison, S., Lyons, T., 2019. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit, in: 2019 Computing in Cardiology (CinC), IEEE. pp. Page–1.

Paoli, C.J., Reynolds, M.A., Sinha, M., Gitlin, M., Crouser, E., 2018. Epidemiology and costs of sepsis in the united states—an analysis based on timing of diagnosis and severity level. Critical care medicine 46, 1889–1897.

Patel, J.J., Taneja, A., Niccum, D., Kumar, G., Jacobs, E., Nanchal, R., 2015. The association of serum bilirubin levels on the outcomes of severe sepsis. Journal of intensive care medicine 30, 23–29.

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

Poggio, T., Liao, Q., 2018. Theory i: Deep networks and the curse of dimensionality. Bulletin of the Polish Academy of Sciences: Technical Sciences .

PyTorch, . https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.

Qu, Z., Ye, Y., Li, F., Ren, Y., Lu, F., Li, L., Lyu, J., Yin, H., 2023. Paco2 levels at admission influence the prognosis of sepsis patients: A nonlinear relationship. Journal of Translational Critical Care Medicine 5, e00012.

Reyna, M.A., Josef, C.S., Jeter, R., Shashikumar, S.P., Westover, M.B., Nemati, S., Clifford, G.D., Sharma, A., 2020. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Critical care medicine 48, 210–217.

Rimmer, E., Garland, A., Kumar, A., Doucette, S., Houston, B.L., Menard, C.E., Leeies, M., Turgeon, A.F., Mahmud, S., Houston, D.S., et al., 2022. White blood cell count trajectory and mortality in septic shock: A historical cohort study. Canadian Journal of Anesthesia/Journal canadien d'anesthésie 69, 1230–1239.

Rodrigues, A.R., Oliveira, A., Vieira, T., Assis, R., Lume, C., Gonçalves-Pereira, J., Fernandes, S.M., 2024. A prolonged intensive care unit stay defines a worse long-term prognosis–insights from the critically ill mortality by age (cimba) study. Australian Critical Care .

Rumbus, Z., Garami, A., 2019. Fever, hypothermia, and mortality in sepsis: Comment on: Rumbus z, matics r, hegyi p, zsiboras c, szabo i, illes a, petervari e, balasko m, marta k, miko a, parniczky a, tenk j, rostas i, solymar m, garami a. fever is associated with reduced, hypothermia with increased mortality in septic patients: a meta-analysis of clinical trials. plos one. 2017; 12 (1): e0170152. doi: 10.1371/journal. pone. 0170152. Temperature 6, 101–103.

Saheed, Y.K., 2023. Effective dimensionality reduction model with machine learning classification for microarray gene expression data, in: Data science for genomics. Elsevier, pp. 153–164.

Santana, A.R., de Sousa, J.L., Amorim, F.F., Menezes, B.M., Araújo, F.V.B., Soares, F.B., Santos, L.C.d.C., de Araújo, M.P.B., Rocha, P.H.G., Júnior, P.N.F., et al., 2013. Sao 2/fio 2 ratio as risk stratification for patients with sepsis. Critical Care 17, 1–59.

Shapley, L.S., et al., 1953. A value for n-person games .

Shashikumar, S.P., Stanley, M.D., Sadiq, I., Li, Q., Holder, A., Clifford, G.D., Nemati, S., 2017. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. Journal of electrocardiology 50, 739–743.

Tautz, T.J., Urwyler, A., Antognini, J.F., Riou, B., 2010. Case scenario: increased end-tidal carbon dioxide: a diagnostic dilemma. The Journal of the American Society of Anesthesiologists 112, 440–446.

Textoris, J., Fouché, L., Wiramus, S., Antonini, F., Tho, S., Martin, C., Leone, M., 2011. High central venous oxygen saturation in the latter stages of septic shock is associated with increased mortality. Critical care 15, 1–6.

Tsantes, A.G., Parastatidou, S., Tsantes, E.A., Bonova, E., Tsante, K.A., Mantzios, P.G., Vaiopoulos, A.G., Tsalas, S., Konstantinidi, A., Houhoula, D., et al., 2023. Sepsis-induced coagulopathy: an update on pathophysiology, biomarkers, and current guidelines. Life 13, 350.

Tuttle, E., Wang, X., Modrykamien, A., 2023. Sepsis mortality and icu length of stay after the implementation of an intensive care team in the emergency department. Internal and Emergency Medicine 18, 1789–1796.

Vilone, G., Longo, L., 2021. Classification of explainable artificial intelligence methods through their output formats. Machine Learning and Knowledge Extraction 3, 615–661.

Wang, Y., Gao, S., Hong, L., Hou, T., Liu, H., Li, M., Yang, S., Zhang, Y., 2023. Prognostic impact of blood urea nitrogen to albumin ratio on patients with sepsis: A retrospective cohort study. Scientific reports 13, 10013.

WHO, 2024. https://www.who.int/news-room/fact-sheets/detail/sepsis.

Yang, J., Hao, S., Huang, J., Chen, T., Liu, R., Zhang, P., Feng, M., He, Y., Xiao, W., Hong, Y., et al., 2023. The application of artificial intelligence in the management of sepsis. Medical Review 3, 369–380.

Ying, J., Wang, Q., Xu, T., Lu, Z., 2021. Diagnostic potential of a gradient boosting-based model for detecting pediatric sepsis. Genomics 113, 874–883.

Zabihi, M., Kiranyaz, S., Gabbouj, M., 2019. Sepsis prediction in intensive care unit using ensemble of xgboost models, in: 2019 Computing in Cardiology (CinC), IEEE. pp. Page–1.

Zayek, M., Bhat, J., Bonner, K., Blake, M., Peevy, K., Jha, O.P., Gulati, R., Bhat, R., 2020. Implementation of a modified neonatal early-onset sepsis calculator in well-baby nursery: a quality improvement study. Pediatric Quality & Safety 5, e330.

Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L.M., Caterino, J., Zhang, P., et al., 2024. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–18.

Machine Learning Algorithms for Early Sepsis Diagnosis: Predictive Power, Explainability and Economic Advantage

Supplement

Jan Frąckowiak
jj.frackowiak2@student.uw.edu.pl

Department of Data Science, Faculty of Economic Sciences, University of Warsaw

S1. Introduction

Following supplement provides additional materials and detailed analyses to give more context to findings of the study and describe them in finer details. It includes a summary of literary references that offer further insights into the use of predictive algorithms in healthcare, with a specific focus on sepsis diagnosis and the associated challenges. Additionally, the supplement describes the specifics of the algorithms, tests, and training procedures used in the study. What is more, an expanded discussion of the results is presented, concentrating on models performance in time and SHAP values correlations.

S1.1. Advancements in AI for Sepsis Diagnosis

Following discussion of literary studies aims to capture the latest innovations in the rapidly evolving landscape of AI-supported sepsis diagnostics. By establishing a broader context, it facilitates a more in-depth examination of the findings presented in this study.

A recent study from the University of Michigan (Kamran et al., 2024) examined the performance of the Epic Sepsis Model, an AI tool designed to detect sepsis early in hospital settings. The study found that the model faces challenges in accurately distinguishing between high- and low-risk patients prior to treatment. Additionally AI methods used in the U.S. often depend on data influenced by clinician suspicion and primarily predict sepsis onset, which means they can fall short in providing early warnings before sepsis criteria are met.

To address this, the study evaluated the AI's discriminative performance throughout the hospitalization process, focusing on its effectiveness relative to the timing of treatment. The findings indicate that the Epic Sepsis Model performs adequately only when clinicians already suspect sepsis. This timing mismatch limits the tool's utility in early detection and intervention.

The research analyzed data from 77,582 hospitalizations, of which sepsis occurred in 3,766 cases (4.9%). The model achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.62 when including predictions made before sepsis criteria were met and occasionally after clinical recognition. However, when predictions made after clinical recognition were excluded, the AUC-ROC dropped to 0.47. This example underscores the need for providing early warnings before clinical suspicion arises.

Another case study on AI in sepsis diagnosis (Zhang et al., 2024) highlights several challenges encountered in practical deployments of diagnosis support systems. The study reports numerous instances where human-AI collaboration in decision-making has been less successful than anticipated.

The study begins by emphasizing the inherent complexity of sepsis diagnosis. It notes that early sepsis diagnosis is a challenging, often under-explored real-world scenario. This process demands that human experts navigate high levels of uncertainty and make critical, time-sensitive decisions based on incomplete information.

The study then details the obstacles users face with current AI implementations. While today's AI systems for medical decision support may perform well on benchmark datasets and in research settings, they often fall short in real-world applications. This discrepancy results in these systems being frequently neglected or rejected by the medical community. Moreover, because clinicians are held accountable for inaccurate diagnoses, doctors generally trust their own judgment more than AI predictions.

The authors of the study argue that current AI decision support systems (DDS) operate within a so-called 'Competition Paradigm'. These systems are designed primarily to predict final decision outcomes, such as sepsis risk scores that indicate whether a patient has sepsis. Medical practitioners state that this approach inadvertently challenges their authority and expertise and positions AI as a competitor in the decision-making process.

In addition to that, physicians often find AI predictions to be delayed, lacking in explainability, and thus not actionable. Given that clinicians are responsible for medical decisions, they approach AI predictions with significant caution.

In response to these challenges, the authors deviced SepsisLab: a new algorithmic support system designed to enhance earlier stages of the medical decision-making process for sepsis diagnosis, rather than providing a blunt final prediction. Proposed SepsisLab operates within a human-AI collaboration framework and focuses on four key steps:

The first stage - *Generating Hypotheses* - involves physi-

cians evaluating sepsis-risk patients using information from electronic health records (EHR) and physical examinations, despite significant uncertainty. SepsisLab assists by providing relevant data and insights, helping to form initial hypotheses about the patient's condition.

Next, in the *Gathering Data* phase, physicians order laboratory tests based on these hypotheses to collect additional information. The system supports this process by suggesting and prioritizing relevant tests according to the evolving clinical context.

During the *Testing Hypotheses* stage, physicians analyze the lab test results and refine or adjust their initial hypotheses. SepsisLab aids this process by correlating test outcomes with potential diagnoses and suggesting new investigation avenues if necessary.

Finally, in the *Making Decisions* phase, physicians use the refined hypotheses to arrive at a diagnosis. SepsisLab provides decision support by summarizing key findings and trends, but the ultimate diagnosis remains the responsibility of the clinicians

Thus, the authors aimed not to replace clinicians but to enhance their decision-making role by providing explainable and actionable insights focused on disease development. SepsisLab is designed as a complex and interactive system capable of suggesting laboratory tests to reduce the uncertainty associated with predicted sepsis scores. By integrating AI in this way, the system supports clinicians with meaningful data and insights, ultimately improving the diagnostic process without diminishing the critical role of human expertise.

Technically, the authors utilized an architecture featuring an LSTM (Long Short-Term Memory) backbone with a variable attention module. This setup allowed the system to produce fixed-length sequences from input data of varying sizes. Additionally, a collection attention module is employed to integrate the LSTM outputs before they are passed to the final fully connected layer.

The model processes inputs such as vital signs and laboratory values, with LSTM states initialized using patient demographics. The resulting model demonstrated superior discriminative power, as indicated by the AUC-ROC on 10% of validation and 10% of test data. This performance was consistent across datasets with both masked (representing partially missing data) and complete lab values.

The LSTM-based architecture was benchmarked against logistic regression, random forest, and gradient boosting trees (GBT) models, showing better results. Additionally, it received positive feedback from a group of medical experts specializing in sepsis diagnosis.

The challenges highlighted by Zhang et al., 2024, particularly concerning clinician responsibility in diagnosis and the actionability of insights provided by AI diagnostic support systems, could potentially be mitigated through solutions proposed in other related studies.

O'Reilly et al., 2023 explore the future potential of AI in sepsis diagnosis, describing the disease as a global burden. Their insights emphasize the importance of responsibility and trust among medical practitioners when using AI for diagnosis. The authors highlight that robust regulatory frameworks for AI in medicine are essential to instill confidence in clinicians to integrate this technology into clinical practice.

They also observe that the United States FDA has recently begun developing such frameworks. For instance, implementing AI systems can be challenging due to uncertainties about when an AI algorithm is sufficiently validated to be integrated into standardized care processes. The European AI Strategy 2021 suggests that AI products should meet general requirements, including clarity of intended purpose, accuracy, and the assurance that the training data is reliable, representative, and adequately utilized.

Apart from regulations supporting implementation of AI based diagnosis systems, authors of the study also address related ethical consideration. While it might be expected that physicians should have a solid understanding of the AI tools they use, it is often neither feasible nor realistic for a single physician to comprehend or be aware of the weighting of variables within every AI tool. Some AI tools incorporate over 40 variables, a number likely to increase as these tools evolve. The greatest challenge to the prospective implementation of AI in medicine may not be logistical or practical but rather the broader ethical questions regarding the extent of agency we delegate to AI, which society must address.

In the context of the actionability of insights, which Zhang et al., 2024 identified as another major deficiency in current AI implementations for sepsis diagnosis, the development and integration of explainability features could prove crucial.

Zhang et al., 2024 incorporated actionability features into the SepsisLab project, creating a recommendation module that presents a ranked list of missing lab values that could reduce uncertainty in predicting sepsis risk. However, more specific explanatory features in an AI system could enable practitioners to better guide and refine their own insights.

In their inquiry, Band et al., 2023 conducted a systematic review of available XAI techniques, featuring 150 articles. They examined various well-known methods such as LIME, SHAP, and notably Layer-wise Relevance Propagation (LRP), a variant of which was used by Lauritsen et al., 2020 - the foundational reference for the main article of this study. LRP, in particular, could address the challenge of explainability even for complex neural network architectures. The review also discussed a range of XAI methods (including UMAP, ANCHOR, CIU, Trace, Grad-CAM, T-SNE, NeuroXAI, and X-CFCMC) specifically in the context of medical imaging.

The authors categorized XAI method performance into two key areas: usability and reliability. Usability refers to how easy and user-friendly an XAI method is, and whether users find it satisfying to use. Reliability pertains to the consistency and accuracy of an XAI method's results. Additional requirements distinguished by authors encompass:

> **Identity:** Identical instances should have identical explanations.
>
> **Separability:** Non-identical instances should not have identical explanations.

Stability: Stability of an explanation is observed when the explanation undergoes minimal changes in response to minor variations in unimportant features of the data instance.

Similarity: Similarity measures the percentage match between the explanation of the original data sample and the explanation with a feature change of around 1%.

All these measures were described in more detail by Gawantka et al., 2024.

According to Band et al., 2023 assessment LIME exhibited the lowest performance regarding identity for both tabular and text data, whereas SHAP and Anchors perform better in this aspect. For stability, SHAP excelled on the drug review text dataset, while Anchors performed best on the side effects text dataset. All interpretability frameworks demonstrated high performance in separability across both tabular and text data. In terms of similarity, SHAP lead for both tabular and text data, followed by Anchors. Furthermore, Anchors had the longest average time to generate explanations, while SHAP provides the quickest results. Overall, SHAP showed the most balanced performance across usability and reliability metrics for both tabular and text data.

Given that the field is still emerging, the authors emphasize the need for more comprehensive and detailed criteria to effectively evaluate and compare existing XAI methods used in diagnostic detection. However, despite acknowledging the limitations, the review successfully demonstrates that numerous explainability measures are already capable of providing actionable insights.

Overall, aside from the ethical and political considerations related to building trust in AI within society, recent research on AI applications in sepsis diagnosis highlights two pressing needs. Firstly, predictive algorithms must be capable of detecting sepis risk in near future. This is crucial for both medical reasons and economic considerations, as early predictions can lead to significant cost savings. Secondly, to enhance the actionability of results, these algorithms should provide greater explainability regarding their mechanisms, enabling users to better understand and interpret their findings.

These two observations will be addressed in this supplement, along with more detailed descriptions of training and assessment procedures applied in the study.

S2. Preparation, training and tests

This section elaborates on the components of the model training pipeline, detailing the decisions made throughout the process and the utilised resources.

The dataset used for this project was sourced from PhysioNet, which provided hourly measurements of 41 features together with sepsis label. These features encompassed vital signs, laboratory values, and demographic information. The data is publicly accessible and organized into two folders containing files in .psv format, with each file representing a patient: the training set A (20,336 files) and training set B (20,000 files).

Each file contains text representations of 41 features collected by hospital systems. 2934 files belong to septic patients. On Figure S1 the density plots of these features values are presented, stratified by the sepsis label. The sepsis label was recoded to mark all timestamps associated with patients who eventually developed sepsis; this was also a necessary step in the data preparation for modeling.

15% of unique patients data was allocated to validation set and 15% to test set, while the remaining 70% was utilised during training. The data split was performed in a stratified manner, to preserve the proportion of patients with disease (15.5% after preprocessing).

This split allowed for training the models and providing with a representable set of data to optimise the probability threshold (validation set was used in this purpose) and test the model. Validation and test data were also utilised for the sake of explainability and economic assessment of models.

In the data preparation phase, several transformations were applied to enhance model performance and address two critical issues: class imbalance and missing data.

The first challenge, class imbalance, was partially mitigated through recoding the target variable, aligned with the goal of the modeling task. This was justified given that predicting disease progression and overall risk during hospitalization is more beneficial than merely diagnosing, and considering the existing concerns regarding AI-based diagnostic systems in this area (Kamran et al., 2024, Zhang et al., 2024).

Specifically, for patients who eventually developed sepsis during hospitalization, all timestamps were assigned a positive class. This adjustment increased the proportion of the positive class observations to 15.5%, compared to the original 7.3% of patients who developed sepsis.

To further address the class imbalance during the training of neural network architectures, the *pos_weight* parameter was set to 8. This adjustment in the binary cross entropy loss function made the model more sensitive to sepsis cases, as false negatives are generally more detrimental than false positives in this context.

The second challenge, dealing with the high missingness rates - where 26 out of 40 features had more than 90% missing data - was addressed using the BayesianImputer provided by the Scikit-Learn library (Scikit-Learn). This method leverages Bayesian inference to enhance the performance of the imputation process.

The advantage of utilising Bayesian regularization is the ability to compare results performance of alternative models basing on component of the inferred posterior function. The authors of the founding article describing the use of Bayesian inference in interpolation (MacKay, 1992) define two stages of Bayesian inference. The first stage is nothing new from the broadly used likelihood maximisation and involves maximising the nominator of the posterior:

$$P(w|D, H_i) = \frac{P(D|w, H_i)P(w|Hi)}{P(D|H_i)} \tag{1}$$

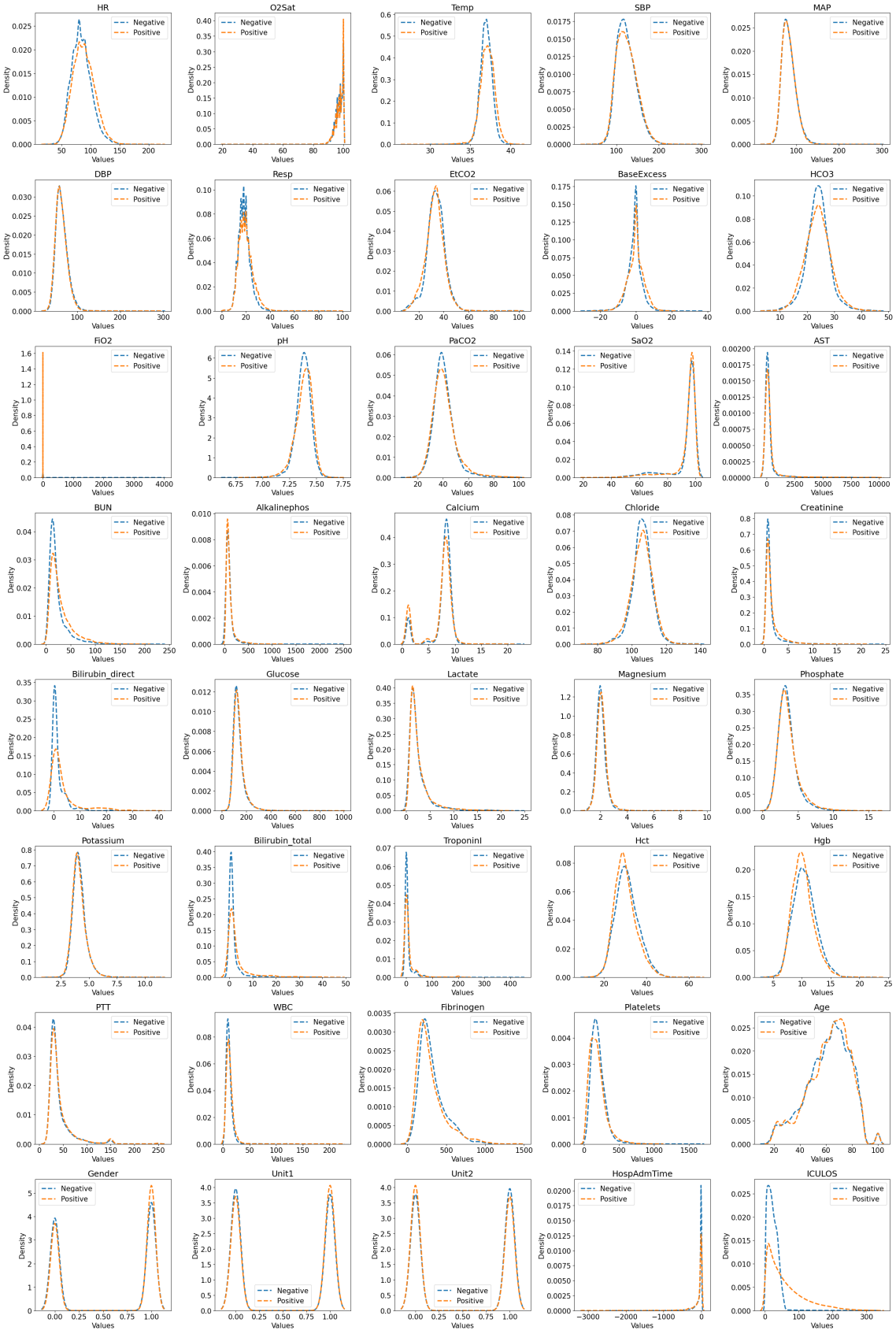


Figure S1: Features Density.
Source: Own preparation.

In words:

$$Posterior = \frac{Likelihood \cdot Prior}{Evidence}$$

where:

w = parameters
D = observed data
H_i = hypothesised theoretical model

However, models chosen with maximum likelihood estimation tend to be overparametrised, and Bayesian inference allows to introduce natural penalty for model's complexity. The second stage of Bayesian inference entails calculating evidence factor for the set of hypothesised models. Evidence, which accounts for the model's complexity and the data, can be approximated by

$$P(D|H_i) \approx P(D|w_{MP}, H_i)P(w_{MP}|H_i)\triangle(w) \quad (2)$$

where:

D = observed data
w_{MP} = parameters of maximal probability
H_i = hypothesised theoretical model
$P(w_{MP}|H_i)\triangle(w)$ = 'Occam's factor'
$\triangle(w)$ = posterior uncertainty in w

Ranking the tested models based on evidence allows for more comprehensive assessment that considers the specific context of the experiment and evaluates the model's ability to generalize effectively. As noted by the authors who proposed this imputation method, the so-called Occam's factor helps penalize model complexity by considering the probability of the parameters given the tested model. While likelihood optimization alone tends to favor overparameterized, complex models, ranking models by their general probability, as measured by evidence, addresses this limitation, promoting models that are both accurate and simple.

As a subsequent step, for variables with low missingness rates, the first differences of variables were computed, and medically justified interactions were calculated. The final preprocessing step before training involved applying Min-Max scaling, since no significant outliers were detected. This scaling method was chosen because it normalizes the data within a defined range of [0, 1], which is beneficial for model performance (de Amorim et al., 2023).

With respect to models development the first step required finding optimal hyperparameters. Three neural net architectures (LSTM, TCN and ANN) were tested with 12, 24 and 12 combinations of hyperparameters respectively. Among these were hidden sizes, numbers of layers, dropout rates, and learning rates. All architectures employed ReLU activation function.

From technical viewpoint, a strategy was employed where input data was read from patient files and calculated on-the-fly. This approach offered high scalability with respect to the number of samples, resulting in low RAM and CPU usage. However, it was quite slow, with the most demanding model (LSTM) taking approximately 8 hours to test all hyperparameter combinations.

To address this, a GPU was utilized, and data was precomputed and stored as a single tensor. This adjustment significantly accelerated training times; however, with larger datasets, it might impact scalability compared to the dynamic strategy. With this approach, training times were reduced to about 1 hour for 12 hyperparameter combinations with LSTM and 1 hour for 24 combinations with TCN, the latter being generally faster to train.

All neural network training was performed using the PyTorch library (PyTorch), while data preparation was handled using basic data operations and Scikit-learn (Scikit-Learn).

Experimentation and hyperparameter optimization were conducted using Optuna and PyTorch Lightning (Lightning), which significantly streamlined the model development process. Optuna (Optuna) was employed for hyperparameter tuning through an efficient grid search, using its Pruner feature to terminate underperforming trials early based on a specified metric - in this case, the AUC-ROC (Area Under the Curve). This approach ensured that only the most promising hyperparameter configurations were explored further, saving computational resources.

PyTorch Lightning enhanced the training workflow by minimizing boilerplate code and organizing the process into a clean, efficient structure. The use of DataModule, ModelWrapper, and Trainer classes in PyTorch Lightning simplified the handling of data, model logic, and training procedures, while also ensuring consistent and organized logging throughout the experiments.

In Figure S2, the architecture of the neural models, configured with the final hyperparameters, is presented. This graph was created using Netron (Netron) application, which enables interactive visualization of models in the .onnx format. The final hyperparameter configurations were selected for each model type: LSTM, TCN, and ANN. All models were trained using a sliding window approach, with a window size of 20 observations. In cases where samples contained fewer than 20 observations, padding was applied to ensure consistency across inputs.

After testing various hyperparameters, all models were trained for 15 epochs on the training data. The values of binary cross-entropy loss and AUC-ROC for both the training and validation sets were monitored and are displayed in Figure S3. The binary cross-entropy loss values are inflated due to the use of the *pos_weight* parameter, which does not allow for direct interpretation of its levels. However, we observe that all models converged smoothly, with loss curves consistently decreasing until the final epoch. For the TCN and LSTM models, the training loss was lower than the validation loss, indicating a tendency to overfit.

Despite this, all models showed an increasing trend in both training and validation AUC-ROC, which was the criterion for early stopping. Consequently, all models reached the predefined maximum of 15 epochs. Additionally, a favorable pattern was observed in the AUC-ROC curves: not only did the metric values improve for both training and validation sets, but the gap between them generally decreased over time.

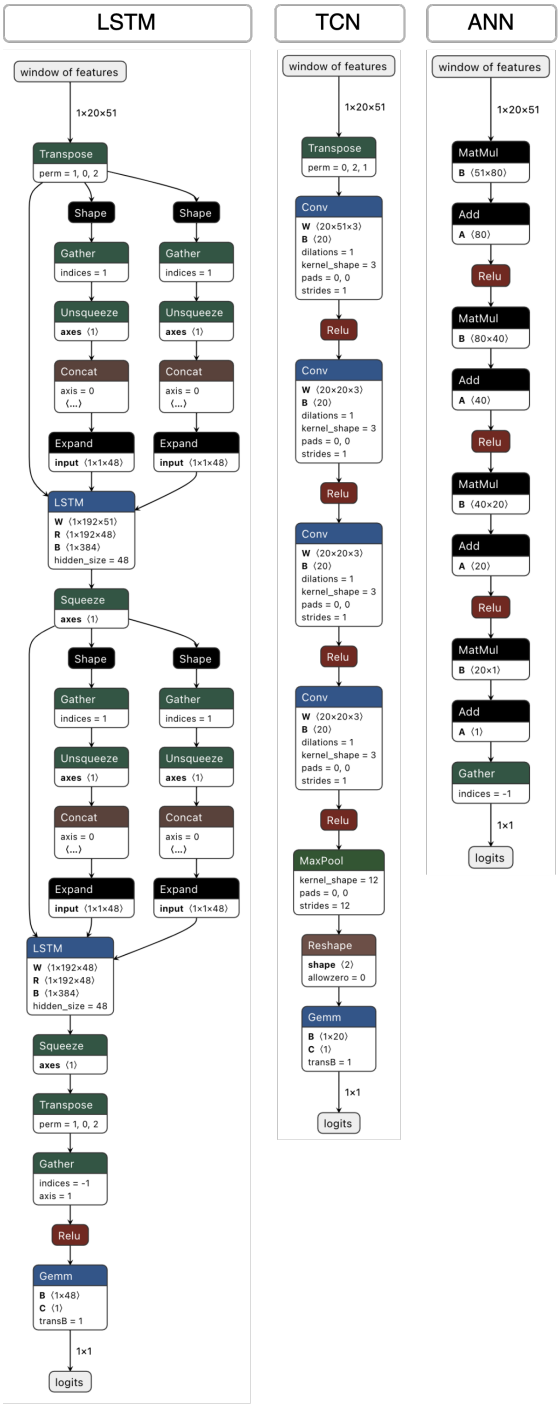


Figure S2: Neural Nets Graphs.
Source: Own preparation with package Netron.

While the models were optimized for discriminative power as indicated by AUC-ROC, binary cross-entropy (BCE) loss was chosen as the loss function for the backpropagation algorithm. Although differentiable proxies for AUC-ROC, such as the implementation available at Github (ROCSTAR), exist, they are not thoroughly tested and do not guarantee better results than the standard BCE function. Moreover, BCE is widely regarded as a reliable representative of AUC-ROC.

GBT architectures - XGBoost, CatBoost, and LightGBM - were fine-tuned using a grid search across 10 distinct hyperparameter combinations. This tuning was performed using 5-fold cross-validation on the training data. The 5-fold cross-validation approach maximizes the information potential of the training data by dividing it into five subsets, where four subsets are used for training and the remaining one for validation. This process is repeated until each subset has been used for validation, and the results are averaged to provide a robust estimate of the model's general performance. GBT models took a single observation of patient data as input.

The Python libraries for XGBoost (XGBoost), LightGBM (LightGBM), and CatBoost (CatBoost) were utilized for model implementation. To ensure reproducibility, a consistent random seed was applied across all experiments.

It's also important to note that the validation and test datasets were prepared in a manner that favored neural network (NN) architectures, which were trained on moving windows of data including 20 most recent observations. To allow for a comparison between the NN and GBT models, the GBT models, which operate on single observations, were evaluated using the last observation from the 20-observation windows used as input for the NN models. As a result, in cases where the data samples were sufficiently long and padding was not required, some of the initial observations were effectively excluded from the GBT model's input.

Had these initial observations been included, it's likely that the GBT models - particularly CatBoost, which demonstrated the highest performance - would have shown even more competitive results.

S3. Models performance in time to onset

CatBoost was found to be the best-performing model, based on a comparative analysis of its performance on validation and test data, as well as an economic assessment against the leading neural network architecture, LSTM. The models were optimized with respect to AUC-ROC - a threshold-invariant metric known for its ability to capture the discriminative power of the model, however the actual accuracy should be evaluated in a more nuanced manner. Specifically, this evaluation should account for the model's ability to detect positive sepsis cases and its capacity to capture the time to clinical diagnosis (sepsis onset).

Time-to-onset was also highlighted by the medical practitioners interviewed in the study conducted by Zhang et al., 2024 as the greatest drawback of existing implementations of diagnostic models, as they often fail to predict sepsis at an earlier stage. Failure to provide early diagnoses was further emphasized as a major shortcoming of existing AI diagnostic decision support systems (AI-DDS) by Kamran et al., 2024.

Sensitivity and specificity are critical metrics in healthcare diagnostics that significantly impact patient care and the effectiveness of medical tests, and could be used to measure models' performance in time.

Sensitivity (also known as true positive rate) refers to a test's ability to correctly identify patients who have the disease. High

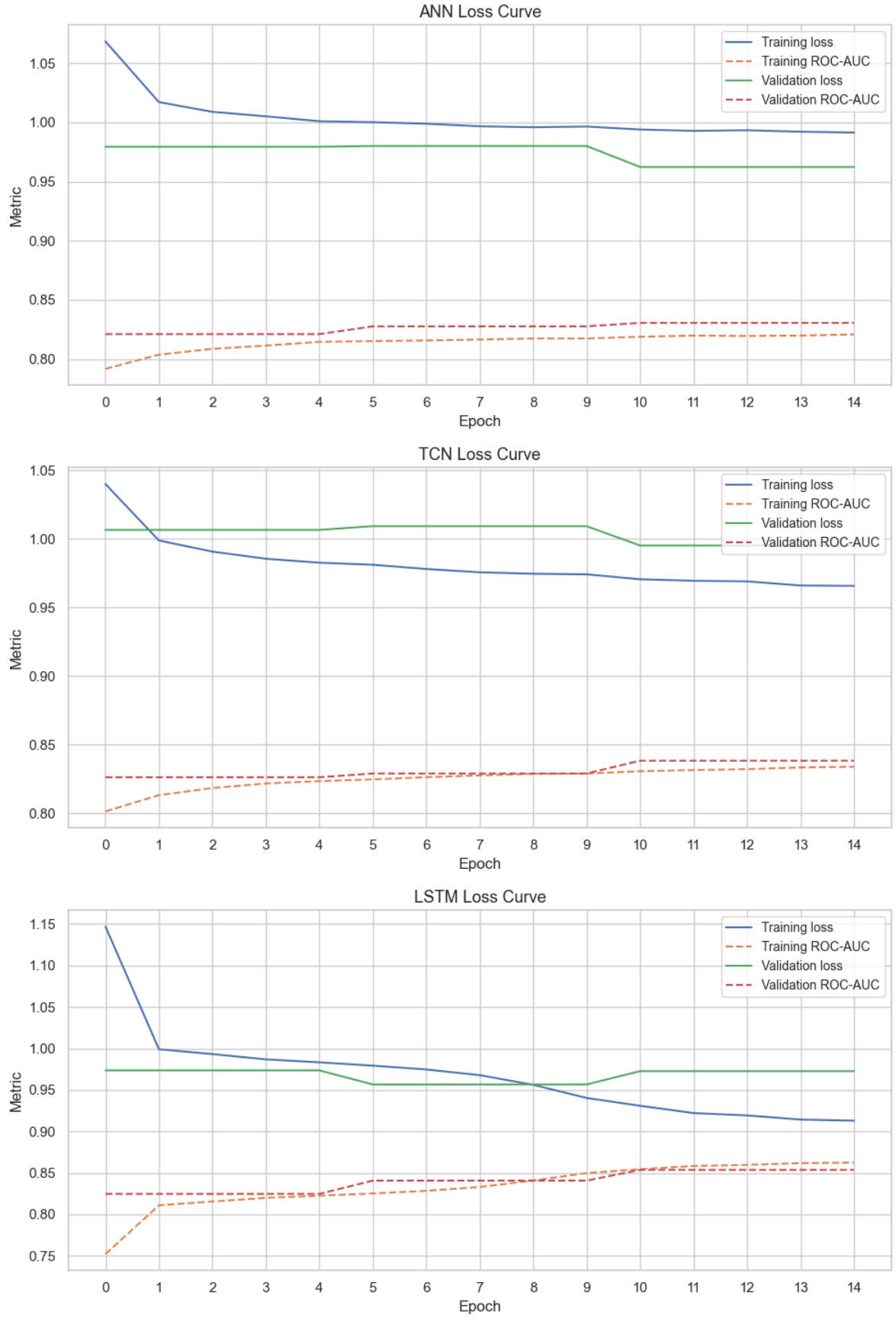


Figure S3: Loss Curves for Neural Net Architectures.
Source: Own preparation.

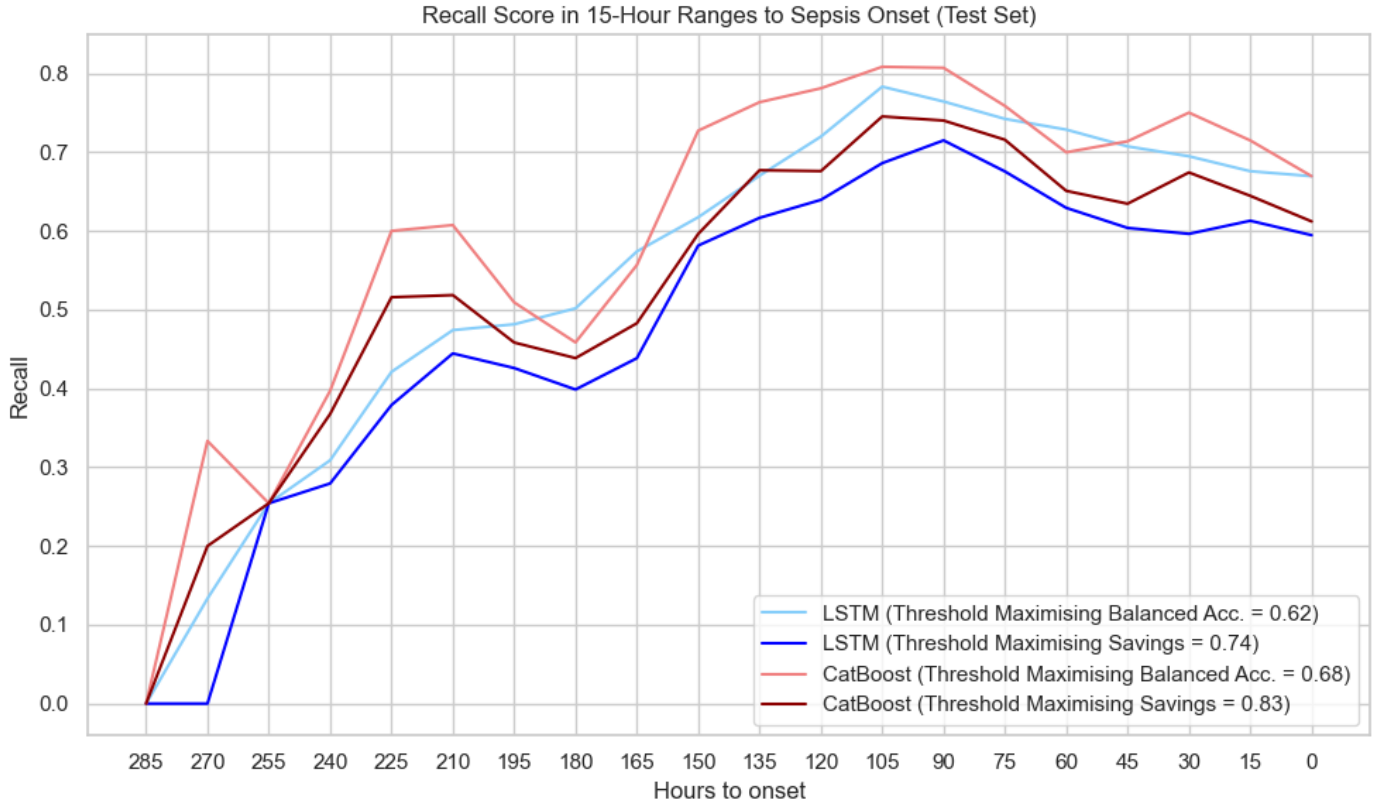


Figure S4: Recall Curves in 15-Hour Ranges on Test Set.
Source: Own preparation.

sensitivity is crucial for early detection and intervention. For example, in cancer screening, a test with high sensitivity ensures that most cases are detected early when treatment options are more effective, thereby improving patient outcomes and potentially saving lives (Curry et al., 2003). In the management of infectious diseases, high sensitivity is essential for promptly identifying infected individuals, which is important for controlling outbreaks and preventing further transmission.

Conversely, specificity measures a test's ability to correctly identify patients who do not have the disease, reflecting the true negative rate. High specificity is important to minimize false positives, which can lead to additional testing, and inappropriate treatments. Balancing sensitivity and specificity is critical for optimizing diagnostic tests.

However, since time-to-onset analysis requires actual illness onset data, specificity could not be directly calculated, particularly for sepsis labels indicating whether a patient developed sepsis during their entire hospital stay. In this context, recall was a more suitable metric for evaluating model performance.

Thus, an analysis of the true positive rate, or recall, was conducted to assess the performance of the top models - LSTM and CatBoost - in identifying positive sepsis cases over time. Recall was chosen for this analysis because it offers valuable insights into the models' ability to detect patients who eventually develop sepsis and thus are more important from medical and economic perspective.

Recall (sensitivity, true positive ratio) is expressed as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

where:

$True\ Positives\ (TP)$ = Correctly predicted positives
$False\ Negatives\ (FN)$ = Misclassified positives

In Figure S4, the recall metric for the test dataset is plotted against the time to sepsis onset, as determined by clinical diagnosis. The performance of CatBoost and LSTM models was evaluated using two different thresholds: one that maximizes relative savings on the validation set (based on assumptions outlined in the main body of the article), and another that maximizes the model's balanced accuracy to visualize recall when the model is tuned solely for performance. The X-axis is divided into 15-hour intervals.

The general trend shows that recall improves as the time to sepsis onset decreases, indicating better model performance as symptoms become more pronounced. It's important to note that, for all models, ICULOS (the time spent in the ICU) was the most influential feature, with a consistently positive effect on sepsis risk. However, the increase in sepsis risk is not linear-it rises until 225 hours before onset, then dips slightly before increasing again after 150 hours. After the 150-hour mark, recall reaches 0.6 and above for thresholds that maximize savings

and 0.7 and above for thresholds that maximize balanced accuracy. This is a considerable lead time before clinical diagnosis.

The thresholds that maximize balanced accuracy (0.62 for LSTM, 0.68 for CatBoost) are generally lower than those optimized for savings (0.74 for LSTM, 0.83 for CatBoost), indicating that they are less restrictive for the minority class.

Except for a brief period around 180 hours before sepsis onset, CatBoost consistently outperforms LSTM at both thresholds. This is notable given that the AUC-ROC for the test set was only slightly higher for LSTM than CatBoost (by 0.002). Despite this, CatBoost is significantly more sensitive to the positive class throughout the entire treatment period. Models that demonstrate substantial recall metrics well in advance of sepsis onset may address the critical need for early diagnosis.

S4. Explainability assessment

The second problem highlighted by Zhang et al., 2024 was the actionability of insights provided by sepsis diagnostic models. Facilitating investigation into factors driving sepsis risk and identifying which laboratory values should be monitored can be addressed through explainability measures. This section offers a more detailed examination of this aspect with regard to conducted study.

For explanation purposes the Dalex (Dalex) and SHAP (SHAP) libraries were utilized, with the concept of Shapley values playing a crucial role in explaining the mechanics of the models. SHAP library was recommended for tabular data in the review of explainability methods prepared by Band et al., 2023.

While the most prominent interrelations between features and model outputs were discussed in the main body of the article, two additional dependency plots demonstrated significant effects of variables and serve as examples of the type of information that could be provided by an explanatory module in an AI-DDS system. These are shown in Figure S5.

First of these illustrates the distribution of the interaction between oxygen saturation and respiratory frequency over the SHAP values for temperature. From this graph, we can observe that as temperature increases, the deviation in its effect on the predicted model output also increases. Fever is often reported in cases of mild or moderate sepsis, but lower body temperatures can also indicate a more severe progression of the disease Rumbus and Garami, 2019. The disparity of effects seen in the model's output reflect this complexity.

In contrast to the varied effects of temperature, a more consistent relationship emerges from the interaction between respiratory rate, oxygen saturation, and temperature. The color progression in the plot highlights a strong negative correlation among these variables.

Within the central temperature range, the model shows high confidence in predicting sepsis risk. Specifically, at a scaled temperature value of approximately 0.8, there is a notable drop in sepsis risk accompanied by a cluster of blue points representing low oxygen saturation and respiratory rate. This drop is associated with the temperature approaching normal values.

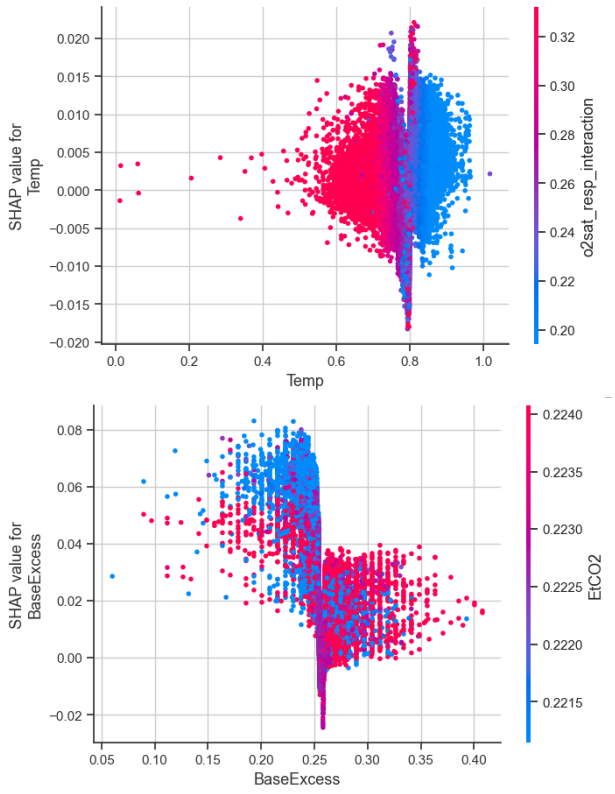


Figure S5: Dependency Plots for CatBoost.
Source: Own preparation.

Subsequently, there is a sharp increase in sepsis risk, marked by a cluster of red points, indicating that a slightly elevated temperature, combined with high oxygen saturation and respiratory rate, is more indicative of sepsis. Beyond this point, the effects become more varied, with lower oxygen saturation and respiratory rate contributing to a more disparate pattern in sepsis risk.

A much stronger relationship with the model output is observed in the second plot on Figure S5, which demonstrates the influence of end-tidal carbon dioxide (ETCO2) on model output alongside the effect of base excess value. In this plot, the maximum effect on the model output reaches up to 0.08, compared to 0.02 in the previous plot.

This plot also underscores the importance and non-linear effect of ETCO2, as discussed in the main article. High SHAP values are observed for both low and high ETCO2 values, while those near the mean value diminish feature's contribution. A positive correlation between base excess and ETCO2 is also evident in the plot. This relationship has been documented in other studies, such as Pishbin et al., 2015, which examined the interdependence between ETCO2 and arterial blood gas parameters and found a significant and strong linear relationship (with $r = 0.346$).

In Figure S6, a correlation heatmap of SHAP values derived from the CatBoost model is presented. This heatmap highlights the correlations between pairs of clinical features, where the absolute value of the correlation coefficient for their SHAP values exceeds 0.33, offering valuable insights into the relationships

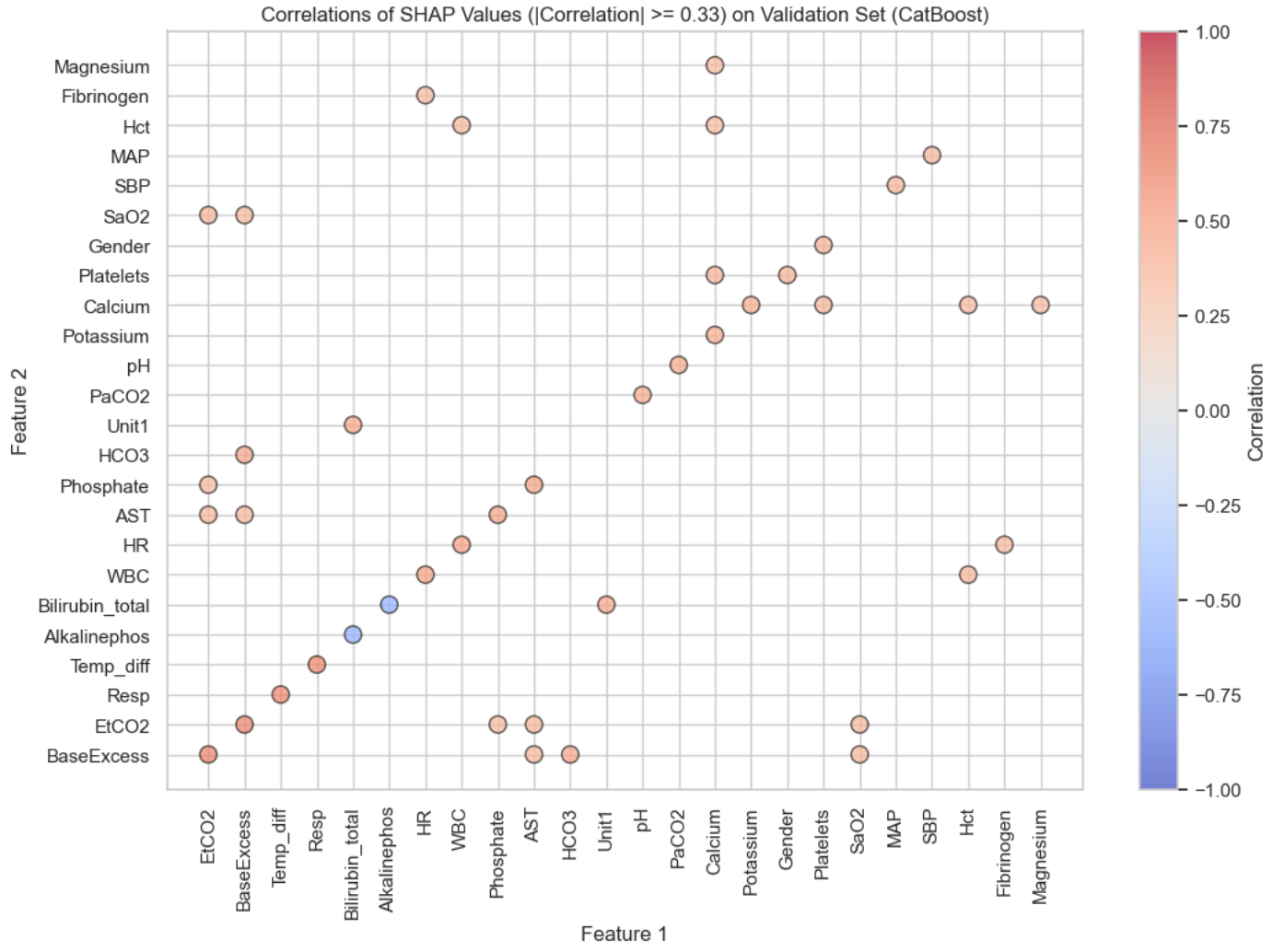


Figure S6: SHAP values correlations for Catboost.
Source: Own preparation.

among different physiological parameters and model output.

The first group of correlations pertains to cardiovascular and inflammatory responses. The positive correlation between systolic blood pressure (SBP) and Mean Arterial Pressure (MAP) is expected since SBP is a major determinant in the calculation of MAP, representing the average arterial pressure during a cardiac cycle (DeMers and Wachs, 2019). Additionally, the correlation between SHAP values for white blood cell count (WBC) and heart rate (HR) reflects the body's inflammatory response during infection or sepsis, where both markers typically increase due to physiological stress. The positive correlation between fibrinogen and HR underscores the interaction between coagulation and cardiovascular response, particularly in inflammatory states.

The second group focuses on respiratory function and acid-base balance. There is a positive correlation between SaO2 (oxygen saturation), EtCO2 (end-tidal CO2), and base excess, highlighting the interrelationship between respiratory efficiency and acid-base homeostasis. Additionally, the correlation between PaCO2 (partial pressure of carbon dioxide) and pH aligns with the respiratory compensation mechanisms in maintaining acid-base balance (Patel et al., 2015).

The third group of correlations pertains to metabolic and electrolyte balance. The effects of magnesium, calcium, and potassium are interrelated, as all three are crucial for maintaining proper electrolyte balance. The correlation between platelets and calcium levels effect reveals a connection between coagulation processes and calcium homeostasis, which is vital for blood clotting. Additionally, the correlation between platelets and gender may indicate physiological differences in coagulation factors between sexes. It is interesting to note that hematocrit (Hct) also shows weak positive correlations with white blood count (WBC) as they had contradicting effect on the risk of sepsis.

Boronat et al. (2017) estimated logistic regression models revealing that serum calcium is associated with anemia. This finding may help explain why hematocrit also shows a positive correlation with calcium, as both are linked to blood health and oxygen-carrying capacity

Liver function is highlighted in the fourth group of correla-

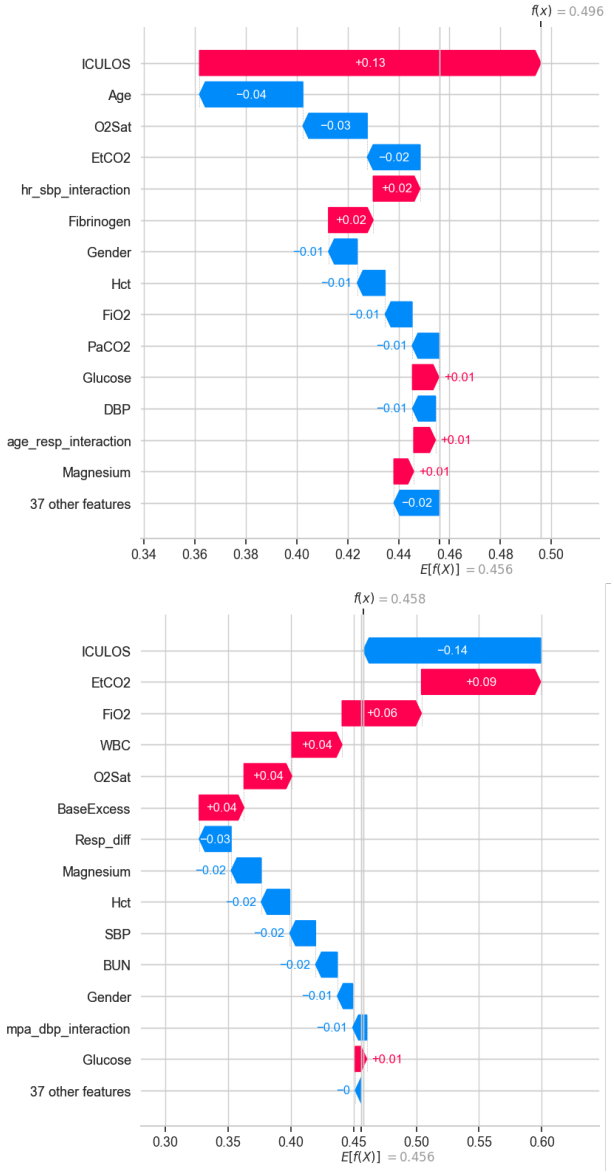


Figure S7: Waterfall Plots for Catboost.
Source: Own preparation.

tions. The positive correlation among base excess, bicarbonate (HCO3), and aspartate aminotransferase (AST) points to liver function, as these markers are indicative of metabolic and hepatic status. Elevated levels of bicarbonate, AST, and base excess typically suggest impaired liver function.

Finally, a notable negative correlation is observed between alkaline phosphatase (Alkalinephos) and bilirubin (Bilirubin_total). According to Wang et al. (2014), bilirubin likely inhibits alkaline phosphatase activity, possibly through negative interference mechanisms.

Overall, the heatmap underscores the complex and multidimensional nature of sepsis, revealing interdependencies between various physiological systems. Analyzing these correlations not only validates the significance of the variables identified in previous analyses but also offers deeper insights into the physiological processes underlying sepsis. Such insights are crucial for enhancing the interpretability of AI-based diagnostic support systems, enabling healthcare providers to better understand model outputs and potentially leading to more timely and accurate clinical interventions.

In Figure S7 waterfall plots are presented. These illustrate the top contributing features for specific instances, represented here as timestamps from patients' records.

Plots decompose the final prediction into a series of contributions from various factors, showing the path the model output takes from the base expected value to the final prediction. Such visual explanations can be instrumental in deriving patient-specific, actionable insights.

Both figures highlight the dominant influence of ICU length of stay (ICULOS), which was also prominently observed in the main analysis. The first plot focuses on the late phase of sepsis, where ICULOS has a significant positive contribution (+0.13). Although the patient was in a later stage of hospitalization, other factors such as laboratory values (O2Sat, ETCO2), gender, and the patient's young age offset this effect, driving the predicted probability down. Without the contribution of ICULOS, the predicted probability would have been as low as 0.36.

In contrast, the second figure depicts an early stage of hospitalization, inferred from the negative contribution of ICULOS (-0.14). Despite the early phase, key clinical markers (ETCO2, FiO2, WBC, O2Sat, BaseExcess) exert a significant upward influence, bringing the predicted risk close to the base sepsis risk.

Such analyses advocate that, beyond explaining the complex interdependence between clinical factors, XAI techniques enable us to recognize specific patient situations and attribute feature effects on a local level. This deeper understanding may lead to insights that could guide clinical interventions

S5. Final Notes

The main article presented a thorough investigation into the potential of AI-based decision support systems (AI-DDS) by focusing on three critical characteristics essential for their adoption: predictive power, explainability, and economic advantage. The study revisited relevant literature, outlined methodologies, and rigorously evaluated these characteristics on a practical example of sepsis diagnosis.

The core of the work detailed the development and evaluation of the models, including the selection process, dataset quality, and applied architectures. It also examined the metrics used to assess explainability and economic benefits, providing both theoretical context and practical implementation insights. The results showed that CatBoost and LSTM models achieved highest levels of predictive power, with CatBoost outperforming in terms of economic savings.

The supplementary section of the article provided additional materials and in-depth analyses to complement the main findings and respond to currently reported challenges. These included a discussion on predictive algorithms in healthcare, with a focus on sepsis diagnosis, as well as more detailed analyses of models' performance over time and the significance of explainability techniques.

In addition to examining the problem of clinical responsibility for diagnosis, a recent study by Zhang et al. (2024) highlighted two ongoing challenges reported by healthcare practitioners working with AI-DDS systems for sepsis. First, the necessity for predictive algorithms to not only forecast the onset of sepsis but also assess its likelihood in the near future. Second, the importance of improving the explainability of AI models to enhance their practical utility. The supplement addressed these insights by further investigating solutions that utilise open-source models.

Presented models demonstrated capability for deeper analysis, enhancing the actionability of results. This was particularly evident in the detailed insights gained from examining the interdependencies between SHAP values for various features, as well as in the analysis of local explanations that could guide interventions in specific patient cases.

Interestingly, CatBoost consistently outperformed the LSTM model, even when evaluated on predicting the onset of sepsis over time - a scenario where the LSTM model, with its sequence-to-one architecture, was theoretically expected to excel. This finding underscores the robustness of the CatBoost model, challenging conventional expectations.

References

de Amorim, L.B., Cavalcanti, G.D., Cruz, R.M., 2023. The choice of scaling technique matters for classification performance. Applied Soft Computing 133, 109924.

Band, S.S., Yarahmadi, A., Hsu, C.C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A.T., Liang, H.W., 2023. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked 40, 101286.

Boronat, M., Santana, A., Bosch, E., Lorenzo, D., Riaño, M., Garcia-Canton, C., 2017. Relationship between anemia and serum concentrations of calcium and phosphorus in advanced non-dialysis-dependent chronic kidney disease. Nephron 135, 97–104.

CatBoost, . `https://catboost.ai/`.

Curry, S.J., Byers, T., Hewitt, M., et al., 2003. Potential of screening to reduce the burden of cancer, in: Fulfilling the Potential of Cancer Prevention and Early Detection. National Academies Press (US).

Dalex, . `https://pypi.org/project/dalex/`.

DeMers, D., Wachs, D., 2019. Physiology, mean arterial pressure .

Gawantka, Just, S., Wappler, L., 2024. A Novel Metric for Evaluating the Stability of XAI Explanations. `https://www.astesj.com/publications/ASTESJ_090113.pdf`. [Online; accessed 18-August-2024].

Kamran, F., Tjandra, D., Heiler, A., Virzi, J., Singh, K., King, J.E., Valley, T.S., Wiens, J., 2024. Evaluation of sepsis prediction models before onset of treatment. NEJM AI 1, AIoa2300032.

Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B., 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications 11, 3852.

LightGBM, . `https://lightgbm.readthedocs.io/en/stable/`.

Lightning, . `https://lightning.ai/docs/pytorch/stable/`.

MacKay, D.J., 1992. Bayesian interpolation. Neural computation 4, 415–447.

Netron, . `https://github.com/lutzroeder/netron`.

Optuna, . `https://optuna.org/`.

O'Reilly, D., McGrath, J., Martin-Loeches, I., 2023. Optimizing artificial intelligence in sepsis management: Opportunities in the present and looking closely to the future. Journal of Intensive Medicine .

Patel, J.J., Taneja, A., Niccum, D., Kumar, G., Jacobs, E., Nanchal, R., 2015. The association of serum bilirubin levels on the outcomes of severe sepsis. Journal of intensive care medicine 30, 23–29.

Pishbin, E., Ahmadi, G.D., Sharifi, M.D., Deloei, M.T., Shamloo, A.S., Reihani, H., 2015. The correlation between end-tidal carbon dioxide and arterial blood gas parameters in patients evaluated for metabolic acid-base disorders. Electronic physician 7, 1095.

PyTorch, . `https://pytorch.org/docs/2.4/`.

ROCSTAR, . `https://github.com/iridiumblue/roc-star`.

Rumbus, Z., Garami, A., 2019. Fever, hypothermia, and mortality in sepsis: Comment on: Rumbus z, matics r, hegyi p, zsiboras c, szabo i, illes a, petervari e, balasko m, marta k, miko a, parniczky a, tenk j, rostas i, solymar m, garami a. fever is associated with reduced, hypothermia with increased mortality in septic patients: a meta-analysis of clinical trials. plos one. 2017; 12 (1): e0170152. doi: 10.1371/journal. pone. 0170152. Temperature 6, 101–103.

Scikit-Learn, . `https://scikit-learn.org/stable/`.

SHAP, . `https://shap.readthedocs.io/en/latest/`.

Wang, Z., Guo, H., Wang, Y., Kong, F., Wang, R., 2014. Interfering effect of bilirubin on the determination of alkaline phosphatase. International Journal of Clinical and Experimental Medicine 7, 4244.

XGBoost, . `https://github.com/dmlc/xgboost`.

Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L.M., Caterino, J., Zhang, P., et al., 2024. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–18.