5-11-2023

Best of Both Worlds: Combining Predictive Power with Interpretable and Explainable Results for Patient Pathway Prediction

Sandra Zilker  
_Friedrich-Alexander-Universität Erlangen-Nürnberg, sandra.zilker@fau.de_

Sven Weinzierl  
_Friedrich-Alexander-Universität Erlangen-Nürnberg, sven.weinzierl@fau.de_

Patrick Zschech  
_Friedrich-Alexander-Universität Erlangen-Nürnberg, patrick.zschech@fau.de_

Mathias Kraus  
_Friedrich-Alexander-Universität Erlangen-Nürnberg, mathias.kraus@fau.de_

Martin Matzner  
_Friedrich-Alexander-Universität Erlangen-Nürnberg, martin.matzner@fau.de_

Follow this and additional works at: _https://aisel.aisnet.org/ecis2023_rp_

Recommended Citation

Zilker, Sandra; Weinzierl, Sven; Zschech, Patrick; Kraus, Mathias; and Matzner, Martin, "Best of Both Worlds: Combining Predictive Power with Interpretable and Explainable Results for Patient Pathway Prediction" (2023). _ECIS 2023 Research Papers_. 266.  
_https://aisel.aisnet.org/ecis2023_rp/266_

This material is brought to you by the ECIS 2023 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2023 Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact _elibrary@aisnet.org_.

BEST OF BOTH WORLDS: COMBINING PREDICTIVE POWER WITH INTERPRETABLE AND EXPLAINABLE RESULTS FOR PATIENT PATHWAY PREDICTION

Research Paper

Sandra Zilker, FAU Erlangen-Nürnberg, Germany, sandra.zilker@fau.de
Sven Weinzierl, FAU Erlangen-Nürnberg, Nürnberg, Germany, sven.weinzierl@fau.de
Patrick Zschech, FAU Erlangen-Nürnberg, Nürnberg, Germany, patrick.zschech@fau.de
Mathias Kraus, FAU Erlangen-Nürnberg, Nürnberg, Germany, mathias.kraus@fau.de
Martin Matzner, FAU Erlangen-Nürnberg, Nürnberg, Germany, martin.matzner@fau.de

Abstract

Proactively analyzing patient pathways can help healthcare providers to anticipate treatment-related risks, detect undesired outcomes, and allocate resources quickly. For this purpose, modern methods from the field of predictive business process monitoring can be applied to create data-driven models that capture patterns from past behavior to provide predictions about running process instances. Recent methods increasingly focus on deep neural networks (DNN) due to their superior prediction performances and their independence from process knowledge. However, DNNs generally have the disadvantage of showing black-box characteristics, which hampers dissemination in critical environments such as healthcare. To this end, we propose the design of HIXPred, a novel artifact combining predictive power with explainable results for patient pathway predictions. We instantiate HIXPred and apply it to a real-life healthcare use case for evaluation and demonstration purposes and conduct interviews with medical experts. Our results confirm high predictive performance while ensuring sufficient interpretability and explainability to provide comprehensible decision support.

Keywords: Patient pathway prediction, predictive business process monitoring, explainable artificial intelligence, interpretable artificial intelligence, deep neural networks, business process management.

1 Introduction

The impact of the pandemic situation highlights the importance of providing high-quality healthcare services. However, to guarantee reliable health protection in times of increasing demand and decreasing resources, it becomes more crucial than ever that the processes of medical institutions run compliant and efficiently in the background. Healthcare processes generally consist of clinical and administrative activities, with medical activities and patient pathways being the focus of analysis and improvement initiatives as they tend to be highly dynamic, complex, ad-hoc, and increasingly multi-disciplinary (Rebuge and Ferreira, 2012). Systematically analyzing patient pathways can help to realize several opportunities, such as improving proactive resource allocation, reducing patients’ waiting times, or identifying potential risks and non-compliant behavior in planned treatments (Mannhardt and Blinde, 2017; Rojas et al., 2016). In the past, patient pathways were mostly analyzed manually, e.g., based on checklists (e.g., Vanhaecht et al., 2006). However, the vast amounts of data ubiquitously generated in healthcare information systems
Hybrid Interpretable and Explainable Predictions

(HIS) provide a vital source to analyze patient pathways in a more automated and, thus, efficient manner. These kinds of methods are subsumed under the term process mining (Mans, Aalst, and Vanwersch, 2015). Of particular interest here are modern approaches from the area of predictive business process monitoring (PBPM) that aim to provide timely information about running process instances to identify risks and issues before or while they develop (Marquez-Chamorro, Resinas, and Ruíz-Cortes, 2018). In this way, organizations can derive recommendations for managing and controlling processes early, which is a promising direction for clinical institutions. Early PBPM approaches chiefly relied on rigid methods that required explicit knowledge about the processes at hand. Nowadays, modern PBPM approaches steadily move towards the application of deep neural networks (DNN) that show superior predictive power while requiring less restrictive assumptions (Harl et al., 2020; Heinrich et al., 2021).

On the downside, DNNs lack transparency because they have a complex internal logic that is not intrinsically understandable (Miller, 2019; Wanner et al., 2020). Such black-box models are not an option, especially in the medical field, where patients’ health might be at risk due to opaqueely derived decisions and actions (Jussupow et al., 2021; Rudin, 2019). These circumstances hamper the dissemination and adoption of powerful PBPM approaches for healthcare environments, resulting in the loss of valuable opportunities for crucial efficiency improvements. As a remedy, we propose a novel artifact that offers predictive analytics capabilities for patient pathways. Specifically, we design a method called Hybrid Interpretable and Explainable Predictions for Patient Pathways (HIXPred) that combines a high degree of model interpretability where possible and explainability where necessary with strong predictive performance, addressing the following research question:

**RQ:** How to design a method that combines predictive power with interpretable and explainable results for patient pathway prediction?

To carry out our research, we follow a design science research (DSR) paradigm. Specifically, we adopt the DSR methodology by Peffers et al. (2007) and structure our work using the DSR publication schema by Gregor and Hevner (2013). Accordingly, our paper is organized as follows: First, we synthesize the background literature and related work, summarize limitations and outline derived design requirements. Next, we provide further details on our applied research approach, followed by a thorough description of our proposed artifact. Subsequently, we instantiate HIXPred and apply it to a real-life event log, evaluate predictive performance, and demonstrate the results’ explainability. Before we conclude the paper, we proceed to discuss the contributions and limitations, and outline future research plans.

## 2 Foundations and Related Work

Our research builds on literature from different streams, such as patient pathways and process analysis, process mining in healthcare, and predictive business process monitoring, which we will elaborate more on in what follows. Based on the literature identified in these sections, we defined limitations and research gaps, which are summarized in Table 1.

### 2.1 Patient Pathways and Process Analysis

Healthcare processes cover all activities related to the diagnosis, treatment, and prevention of diseases to improve a person’s well-being (Mans, Aalst, and Vanwersch, 2015). This includes patient-related activities organized in patient pathways and administrative activities supporting clinical tasks (Rebuge and Ferreira, 2012). Patient pathways are directly linked to a patient’s diagnostic-therapeutic cycle and therefore do not constitute strictly standardized processes. Instead, they are planned procedural patterns based on medical guidelines that are tailored to individual circumstances (e.g., a patient’s condition or availability of resources) (Lenz and Reichert, 2007). If restricted to a single clinical setting, we may also refer to clinical pathways, while it is also increasingly common to look beyond inter-organizational borders focusing on so-called patient pathways from a more comprehensive viewpoint (Richter and Schlieter, 2019).
Hybrid Interpretable and Explainable Predictions

Medical processes are typically performed in environments under continuous change and require multi-disciplinary resources (e.g., nurses, physicians, or technical specialists). Thus, unlike other domains, there are ample opportunities for process improvements due to their highly dynamic and complex nature, which makes them attractive for systematic analysis initiatives (Mans, Aalst, and Vanwersch, 2015). Traditionally, processes were analyzed manually, using checklists, interview techniques, document analyses, and conceptual process modeling languages (e.g., Burwitz, Schlieter, and Esswein, 2013; Vanhaecht et al., 2006). However, such approaches reach their limits as they are time-consuming, static, and prone to errors due to subjective biases (Rebuge and Ferreira, 2012). Consequently, healthcare environments require more flexible, efficient, and fact-based analysis approaches by exploiting digital trace data that are vastly generated in different types of involved HIS. Such systems usually store hundreds of tables with patient-related event data that can be used to improve patient pathways while reducing unnecessary costs (Dallagassa et al., 2021; Mans, Aalst, and Vanwersch, 2015; Rojas et al., 2016).

2.2 Process Mining in Healthcare

The field of process mining offers various techniques for the extraction of process-related knowledge from digital trace data or event logs that are readily available in today’s process-aware information systems (van der Aalst, 2016). On this basis, it is possible, for example, to automatically discover process models from recorded events, identify non-compliant behavior, improve workflows based on detected bottlenecks, monitor running operations in real time, or predict performance deviations before they occur (van der Aalst, 2016). Because of the benefits of such data-driven approaches, we can observe a considerable growth of process mining initiatives in recent years in the healthcare domain (e.g., Dallagassa et al., 2021; Mans, Aalst, and Vanwersch, 2015; Rojas et al., 2016). Here, the focus is primarily on the automated discovery of process models for a better understanding and a reconstruction of patient pathways, followed by initiatives concerned with the analysis and evaluation of resources and the validation of process conformance (Dallagassa et al., 2021). Nevertheless, most process analysis endeavors in healthcare settings are concerned with post-mortem analysis. That is, pathways are examined reactively after they have already been completed. At this point, there is still great potential for improvement through the development and dissemination of more forward-looking, proactive approaches that are capable of analyzing patient pathways during their execution and generating insights about their further course, which is generally addressed in the field of PBPM (Marquez-Chamorro, Resinas, and Ruiz-Cortes, 2018).

2.3 Predictive Business Process Monitoring and Deep Learning

PBPM has emerged in recent years as a sub-field of process mining concerned with the central question of how an ongoing case or a path will unfold based on its past record (Marquez-Chamorro, Resinas, and Ruiz-Cortes, 2018). This may involve the prediction of the next activity or a sequence of activities, the remaining cycle time, or the outcome of a running process instance, which generally helps involved stakeholders anticipate deviating behavior and risks before they occur (Di Francescomarino et al., 2018). The basis of such predictions are methods and models from the fields of statistics and machine learning (ML) (Bishop, 2006). Early PBPM approaches like probabilistic finite automatons, and hidden Markov models chiefly relied on rigid assumptions and explicit process representations (e.g., Breuker et al., 2016; Lakshmanan et al., 2015). That is, they required explicit ex-ante knowledge about the form of an underlying process model (Marquez-Chamorro, Resinas, and Ruiz-Cortes, 2018), which is sometimes challenging to discover or reconstruct from complex event logs. Especially in healthcare environments, which are subject to a high degree of complexity and dynamism, the reconstruction of process models is highly challenging and requires much effort, as exemplified by Mannhardt and Blinde (2017).

Newer approaches based on artificial neural networks that are organized in multi-layered network architectures overcome this limitation. They can discover intricate structures in vast amounts of data and learn complex patterns relevant to a prediction task, commonly known as deep learning (DL) (Janiesch,
Hybrid Interpretable and Explainable Predictions

Zschech, and Heinrich, 2021; LeCun, Bengio, and Hinton, 2015). DNNs typically do not depend on any ex-ante assumptions about a process model and offer superior predictive performance outperforming traditional PBPM approaches (e.g., Kratsch et al., 2020; Rama-Maneiro, Vidal, and Lama, 2021; Wang et al., 2019). This makes them particularly interesting for complex and dynamic process environments.

However, the high performance of DNNs comes at a certain cost as their nested, multi-layered structure creates a lack of transparency by constructing internal high-degree interactions between input features, which are not intrinsically explainable. In other words, it is not clear what information in the input data drives the DNN models to generate their decisions. Therefore, they are typically regarded as black boxes, and further methods or adaptions are required to provide explainability to end users (Janiesch, Zschech, and Heinrich, 2021). Such methods are increasingly discussed under the term explainable artificial intelligence (XAI) (Bauer, Zahn, and Hinz, 2023; Senoner, Netland, and Feuerriegel, 2022).

2.4 Interpretable and Explainable Predictive Business Process Monitoring

Interpretability refers to ML models where the decision logic of the model itself is transparent, such as in decision trees and linear/logistic regressions (Ciocan and Mišić, 2022; Zschech et al., 2022). In interpretable models, humans can tell the output of the model for an unseen input. This allows, for instance, to detect the flawed behavior of the model for input values that are not represented in the training data. In contrast, explainability refers to methods that allow the explanation of the model output for a given input (Du, Liu, and Hu, 2019). However, an explanation is post-hoc, and thus, for an unseen input, it is very difficult to tell the output without letting the model compute it (Rudin, 2019). Consequently, interpretable models are a sub-class of explainable models with drastically increased transparency.

Multiple approaches have been proposed for different prediction targets concerning interpretable models in PBPM. For example, Breuker et al. (2016) proposed a probabilistic-based approach, Böhmer and Rinderle-Ma (2020) a rule-based approach, and Senderovich et al. (2017) a regression-based approach. However, these approaches lack the ability to explicitly model the sequence in which activities appeared. This limits predictive performance, which is crucial in the context of healthcare.

Other researchers have focused on XAI techniques for PBPM (e.g., Rehse, Mehdiyev, and Fettke, 2019) to overcome missing explainability, limiting the applicability of PBPM models (Marquez-Chamorro, Resinas, and Ruiz-Cortes, 2018). These works can be classified into two streams of research (Stierle et al., 2021). One stream uses surrogate models to approximate the model’s behavior locally (e.g., Mehdiyev and Fettke, 2020). Other streams investigate the weights of neural networks (e.g., Evermann, Rehse, and Fettke, 2017; Harl et al., 2020; Rehse, Mehdiyev, and Fettke, 2019; Sindhgatta et al., 2020), the dynamics of sequential models (Evermann, Rehse, and Fettke, 2017), or the gradients (Weinzierl et al., 2020).

3 Research Method

We chose a DSR approach for our research project as it constitutes a fundamental paradigm in IS research that is concerned with the construction of socio-technical artifacts to solve organizational problems and derive prescriptive design knowledge (Gregor and Hevner, 2013). Specifically, we followed the DSR methodology by Peffers et al. (2007) consisting of six phases: (i) problem identification and motivation, (ii) definition of the objectives for a solution, (iii) design and development, (iv) demonstration, (v) evaluation, and (vi) communication. The adoption of the methodology to our project is depicted in Figure 1.

In the first phase, we identified challenges in healthcare environments that affect the analysis and improvement of patient pathways. We looked into the broad body of knowledge using academic literature to consider existing solution approaches and examined their limitations. In particular, we considered the

---

1 We distinguish between the terms “interpretation” and “explanation”. Interpretation is derived from models designed to be intrinsically interpretable, whereas an explanation is created by applying a post-hoc XAI approach to a black-box model.
three research streams (i) patient pathway analysis, (ii) process mining in healthcare, and (iii) predictive 
business process monitoring. Especially review papers, such as those provided by Dallagassa et al. (2021), 
Marquez-Chamorro, Resinas, and Ruiz-Cortes (2018), Rama-Maneiro, Vidal, and Lama (2021), Richter 
and Schlieter (2019), Rojas et al. (2016), and Vanhaecht et al. (2006), helped us to gather a broad overview 
of the existing knowledge base, to examine limitations and research gaps, and to derive requirements 
for the design of a novel artifact. A summary is provided in Table 1, and further details are outlined in 
Section 2.

<table>
<thead>
<tr>
<th>Research Areas</th>
<th>Example References</th>
<th>Limitations/ Research Gaps</th>
<th>Derived Design Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis of healthcare processes/ patient pathways</td>
<td>Richter and Schlieter (2019) and Vanhaecht et al. (2006)</td>
<td>Manual analyses of pathways are static, time-consuming, and prone to errors due to subjective biases.</td>
<td>R1: The method should exploit digital trace data (i.e., recorded events) and support fully automated analyses.</td>
</tr>
<tr>
<td>Process mining in healthcare</td>
<td>Dallagassa et al. (2021), Mans, Aalst, and Vanwersch (2015), Rebuge and Ferreira (2012), and Rojas et al. (2016)</td>
<td>Existing process mining initiatives primarily focus on reactive, post-mortem analyses.</td>
<td>R2: The method should provide proactive analysis capabilities (i.e., predictions) during the execution of patient pathways.</td>
</tr>
<tr>
<td>Process-aware predictive business process monitoring</td>
<td>Breuker et al. (2016), Lakshmanan et al. (2015), and Marquez-Chamorro, Resinas, and Ruiz-Cortes (2018)</td>
<td>Traditional PBPM approaches require explicit knowledge about the structure of process models.</td>
<td>R3: The method should be flexible enough to work without ex-ante assumptions about the process model.</td>
</tr>
<tr>
<td>Predictive business process monitoring with deep learning</td>
<td>Kratsch et al. (2020) and Rama-Maneiro, Vidal, and Lama (2021)</td>
<td>DNNs generally constitute black-box models as their internal decision logic is not intrinsically explainable.</td>
<td>R4: The method should provide high predictive performance while retaining a high degree of interpretability and explainability of the results.</td>
</tr>
</tbody>
</table>

Table 1. Synthesis of background literature.

In the third phase, we designed our novel artifact tailored for patient-related healthcare processes to provide 
interpretable and explainable predictive analysis capabilities in patient pathways. For demonstration 
purposes, we instantiated HIXPred and applied it to a healthcare use case with real-life event log data. 
Specifically, we demonstrate HIXPred’s interpretation and explanation capability. Further, we evaluated the 
predictive performance with different metrics and benchmarked HIXPred to several baseline approaches. 
Finally, the conference publication aids in communicating our results. Since the DSR methodology is an 
iterative process (Peffers et al., 2007), we will use this conference format to get feedback for the next 
itration cycles, where we aim to apply HIXPred in practice.

4 Artifact Description

This section describes the design of our novel artifact that predicts the outcomes of patient pathways during 
run-time with high predictive performance and provides meaningful interpretations and explanations of
these predictions to healthcare providers. HIXPred is designed to fulfill the four requirements derived from the knowledge base (cf. Table 1). In line with R2 and previous research on PBPM (Marquez-Chamorro, Resinas, and Ruiz-Cortes, 2018), HIXPred is structured into an offline and online component, as shown in Figure 2. With respect to R1, the offline component first receives as input an event log documenting patient pathways of a healthcare process and transforms it into sequential patient data, static patient data, and label data. Second, the offline component learns a DNN model based on that data. Sequential and static patient data are used as the source for learning the DNN model, whereas label data describe the learning target, i.e., a process outcome. As such, HIXPred does not depend on any process models, as specified by R3. Third, the offline component creates interpretations and explanations in the form of visual outputs, fulfilling R4. The output with interpretations is a coefficient plot, visualizing coefficient values for static patient data attributes extracted from the DNN model’s last layer. The output with explanations is a SHapley Additive exPlanations (SHAP) plot, visualizing the effect of sequential patient data attributes on the DNN model’s process outcome prediction. The online component receives as input a running process instance and transforms it into static and sequential patient data. Subsequently, the component applies the DNN model learned in the offline component to predict the process outcome. Lastly, the online component presents interpretations and explanations to process stakeholders, i.e., the coefficient and SHAP plot, allowing them to interpret and explain, respectively, the outcome prediction calculated by the DNN model. In the following sections, we detail the six phases of both components.

![Figure 2. Offline and online component of HIXPred.](image)

### 4.1 Transform Event Log

The offline component receives as input an event log \( L \) documenting patient pathways of a healthcare process and transforms it into static patient data \( X_{\text{static}} \), sequential patient data \( X_{\text{seq}} \), and label data \( y \). The transformation procedure consists of three phases.

First, an event log \( L \), that is a set of traces, is loaded. A trace \( \sigma \) (a patient pathway in our case) refers to a process instance and represents a sequence of events, \( e_1, \ldots, e_T \), with sequence length \( T \). An event \( e \) is a tuple \( (p, a, t, d_1, d_2, \ldots, d_n) \), where \( p \) is the id of the process instance, \( a \) is the process activity, \( t \) is the timestamp, and, for example, \( d_n \) is the \( n \)-th data attribute belonging to this event.\(^2\) A data attribute has either the same value across all events of a trace and stores, in our case, static patient data, or it has a different value per event and stores sequential patient data.

Second, static patient data \( X_{\text{static}} \) and sequential patient data \( X_{\text{seq}} \) are created based on the event log \( L \). \( X_{\text{static}} \) lies in \( R^{s \times u} \), where \( s \) denotes the number of prefixes of the event log \( L \)'s traces, and \( u \) is the number of static patient attributes. A prefix is a sub-sequence of a trace or the complete trace itself. Here, prefixes are created from traces to learn a DNN model, which can predict outcomes per time step in running healthcare processes. As \( X_{\text{static}} \) only stores static patient data, prefixes of a trace store the same patient data. Attributes of \( X_{\text{static}} \) can be either binary, categorical, or numerical. Values of a binary attribute are mapped to 0 or 1, categorical values are one-hot encoded, and numerical values are scaled into the range \([0, 1]\). \( X_{\text{seq}} \) lies in \( R^{q \times 1} \), where \( q \) is the vector size of the event log’s

\(^2\) For simplicity, we neglect the index of an event within the sequence in the following.
Hybrid Interpretable and Explainable Predictions

longest prefix, and \( r \) refers to the number of unique process activities. The remaining space is padded with zeros for traces that are shorter than the longest prefix. All process activities are encoded by custom one-hot vectors of length \( r \). A process activity \( i \) is encoded by a vector containing only zeros, except for the \( i \)-th position. If there is no numerical information returned from the \( i \)-th process activity (e.g., from lab tests), we represent the activity by 1, else we represent it by its returned value. The values are then scaled between 0 and 1. Activities are represented by this customized one-hot encoding to model the relationship between activities and their existing numerical attributes explicitly in the data. An example of such a pair could be the activity “measure high blood pressure” and the numerical attribute “blood pressure value”.

Third, label data \( y \) is created for a prediction target based on the event log \( L \). A prediction target describes the process outcome of interest. More specifically, the appearance of an activity of interest within a trace (i.e., \( y \) has the value 1 for all events of this trace) or the absence of the activity (i.e., \( y \) has the value 0 for all events of this trace). The labels are set on a trace level because we strive to predict process outcomes as early as possible, as is common in outcome-oriented PBPM (Teinemaa et al., 2019). For example, if the activity of interest is “admission to intensive care unit (ICU)”, it is predicted at each time step of running process instances if this activity will appear.

Finally, we discard all prefixes and corresponding labels where the activity of interest is part of the sequential data. This is important to avoid data leakage problems in patient pathway predictions.

4.2 Learn DNN Model

For given label data \( y \), static patient data \( X_{\text{static}} \), and sequential patient data \( X_{\text{seq}} \), a DNN model \( \mathcal{M} \) is learned as follows. Let \( X_{\text{static}} \) and \( X_{\text{seq}} \) denote our (training) data set with the corresponding label data \( y = (y_1, \ldots, y_s) \). For a given prefix \( i \in \{1, \ldots, s\} \), a single prefix comprises two sources of data: static patient data \( X_{i,\text{static}} \in \mathbb{R}^{1 \times n} \) and sequential patient data \( X_{i,\text{seq}} \in \mathbb{R}^{1 \times q \times r} \).

Our proposed DL architecture, which naturally models the event log data through two modules, and, at the same time, keeps a high level of interpretability, is illustrated in Figure 3. The first module uses a bi-directional long short-term memory (Bi-LSTM) layer (Graves, Jaitly, and Mohamed, 2013) to encode the patient’s sequential data \( X_{i,\text{seq}} \) into a vector \( h_i \in \mathbb{R}^{m \times 1} \), where \( m \) denotes the number of neurons in the Bi-LSTM. This vector \( h_i \) is usually referred to as the hidden state in a Bi-LSTM. The second module uses a simple feed-forward layer with sigmoid activation (equivalent to logistic regression) to map both the static patient data \( X_{i,\text{static}} \) but also the vector \( h_i \) onto an estimated probability of appearance of the target activity \( \hat{y}_i \). This is achieved by concatenating the patient data \( X_{i,\text{static}} \) with the vector \( h_i \). All parameters from both modules are combined into one DNN model in which the parameters are trained simultaneously. The training objective is to minimize the loss between the estimated appearance of the target activity \( \hat{y}_i \), and the true label \( y_i \), denoting the actual appearance of the target activity.

Mathematically, given parameters \( \beta_{\text{static}} \) and \( \beta_{\text{seq}} \) for the feed-forward and the Bi-LSTM module, respectively, we estimate \( y_i \) as follows:

\[
\hat{y}_i = \frac{1}{1 + \exp(-x_i^T \beta_{\text{static}})}, \tag{1}
\]

\[
x_i^T = (X_{i,\text{static}}, h_i), \tag{2}
\]

\[
h_i = \text{Bi-LSTM}(X_{i,\text{seq}}, \beta_{\text{seq}}), \tag{3}
\]

with \((X_{i,\text{static}}, h_i)\), denoting a concatenation of the static patient data \( X_{i,\text{static}} \), and the Bi-LSTM’s hidden state \( h_i \).

---

3 We consider the case of binary classification. A generalization to multiple classes is feasible using a softmax activation.
4.3 Retrieve Interpretations and Create Explanations

For our learned DNN model, interpretations are retrieved, and explanations are created. Regarding interpretations, the DNN model is designed so that static patient attributes only affect the outcome linearly (see Eq. (1)). Consequently, for a static attribute, the corresponding coefficient in $\beta_{\text{static}}$ denotes the impact of the attribute on the outcome. This allows for the full interpretability of static attributes. The information on coefficients is summarized in a plot $E_{\text{coeff}}$. Concerning explanations, the method builds on the latest research in explaining DNN models using SHAP values (Lundberg and Lee, 2017). SHAP is based on the game theoretically optimal Shapley Values and is a method to explain individual predictions. The goal of SHAP is to explain the prediction of a prefix by computing the contribution of each attribute to the prediction. SHAP comes with two benefits. First, it allows for assessing the impact of an attribute on the estimation for non-linear (potentially sequential) models. As we later show, this is crucial in our case since the Bi-LSTM increases the prediction performance considerably. Second, SHAP explains predictions on an individual level, that is, for each prefix. As we use a custom one-hot-encoding with numerical values, this allows for providing non-linear relations between process activities of prefixes and process outcome predictions (in our case, the occurrence of an activity in prefixes). Finally, information on SHAP values is summarized in a plot $E_{\text{shap}}$.

4.4 Transform Running Process Instance

The online component receives as input a running process instance $\sigma$, that is a prefix of a trace, and transforms it into static patient data $X_{\text{static}}$ and sequential patient data $X_{\text{seq}}$. Even though the transformation procedure for the running process instance $\sigma$ seems similar to the event log $L$’s procedure, it differs in two ways. First, it does not retrieve label data because the DNN model will predict it based on $X_{\text{static}}$ and $X_{\text{seq}}$. Second, $X_{\text{static}}$ and $X_{\text{seq}}$ lay in $\mathbb{R}^{1 \times u}$ and $\mathbb{R}^{1 \times q \times r}$ because they only hold one prefix, representing a running process instance.

4.5 Apply DNN Model

After transforming a running process instance $\sigma$ into static patient data $X_{\text{static}}$ and sequential patient data $X_{\text{seq}}$, our DNN model can be applied, as described by Eqs. (1)–(3). For a given prefix $i$, we use the trained Bi-LSTM module to map the sequential patient data to a vector $\mathbf{h}_i$. Second, we concatenate the static patient data $X_{i,\text{static}}$ with the previously computed hidden state vector $\mathbf{h}_i$. Finally, we multiply the concatenated vectors with the coefficient vector $\beta_{\text{static}}$ and apply the sigmoid function. The resulting value denotes the estimated probability of an activity happening in the future $\hat{y}_i$ (e.g., admission to ICU). If the probability $\hat{y}_i$ is $\geq 0.5$, the activity is predicted ($\hat{y}_i = 1$); otherwise, it is not predicted ($\hat{y}_i = 0$).

4.6 Present Interpretations and Explanations

Lastly, the prediction for a given prefix $\sigma$ is equipped with an interpretation and explanation in the form of a visual output. First, the coefficients in $\beta_{\text{static}}$ are presented that correspond to the static attributes using a
bar chart \( (E_{\text{coef}}) \). Thereby, healthcare providers can understand the DNN model’s general reasoning and also monitor the DNN model’s estimation for a single prefix. Second, the SHAP values are presented for a set of prefixes from the sequential patient data. For this, a swarm plot is used \( (E_{\text{shap}}) \), where the x-axis shows the SHAP values for the respective attributes, and the color denotes the value of the attributes.

5 Evaluation and Demonstration

To profoundly evaluate and demonstrate HIXPred, we instantiated HIXPred and applied it to a real-life healthcare use case. The use case considers a low-level perspective, focusing on the individual procedures of sepsis patients who had undergone, e.g., a treatment or laboratory analysis. Additionally, based on HIXPred’s output, we conducted interviews with two medical experts.

5.1 Use Case Description

The use case is from a Dutch hospital with approximately 50,000 patients per year and 700 beds (Mannhardt and Blinde, 2017). The data set\(^4\) focuses on sepsis cases. Sepsis can be life-threatening and occurs when the body reacts to an infection, thereby injuring organs or tissues. To track all performed patient events, the hospital uses an ERP system. The process consists of logistical activities, including the patient’s stations through the hospital, and patient activities, such as the blood values that were measured and the treatments that were performed. Although the aforementioned process can be described in a fairly structured manner based on the information provided by the use case provider, this structure is reflected to a limited extent in the underlying patient event log. In addition, patients can perform different activities in highly individual pathways, making it difficult to manually detect patterns for the estimation of an individual patient pathway’s outcome.

To ensure sufficient data for the predictions and limit computational time, we only considered patient pathways that were longer than two but shorter than 100 patient events. We also removed patient pathways that did not start with activity ER registration because we assumed this activity to be the central entry point into the patient pathway. The patient event log representing the use case contained 992 patient pathways with 788 different variants over a period of 1.5 years. Patient events can be differentiated into 16 activities with different purposes, for example, release type, measurement of CRP, Leucocytes, and LacticAcid, or stating whether the patient was admitted to ICU or normal care unit, which represent sequential features.

In addition to the control-flow information, the patient event log contained 26 features. Three of them are sequential and numerical and represent the measured values of CRP, Leucocytes, and LacticAcid. Furthermore, patient age is a numerical but static feature. In addition to age, there are 22 additional static features that are either binary or categorical (Mannhardt and Blinde, 2017).

For our evaluation, we consider the prediction target “admission to ICU”, which constitutes a possible outcome of a patient’s pathway, indicating whether a patient is required to transfer to the ICU.

5.2 Experimental Setup

The experimental setup comprises (i) model evaluation, (ii) model selection, and (iii) used baselines. First, in the model evaluation, we split the data into training, validation, and testing sets. To consider changes over time in the analyzed processes, we used the first 55% of the patient pathways for training, the following 15% for validation, and the last 30% for testing. In this process, we retain chronological order within and across the prefixes. Then, we selected the best model through a grid search using training and validation sets and applied the best model to the test set to compute test performance. As a selection criterion for the grid search, we calculated the \( \text{AUC}_{\text{ROC}} \). To measure the performance of the ML models on the test set, we calculated support, precision, recall, and F1-score class-wise, that is, for

---

\(^4\) https://data.4tu.nl/articles/dataset/Sepsis_Cases_-_Event_Log/12707639
the positive and negative class, and the $AUC_{ROC}$ across both classes. We repeated training and testing ten times and calculated the average over these executions, as well as the standard deviation. Moreover, we selected the model with the highest $AUC_{ROC}$ from the ten executions to retrieve interpretations and create explanations. As baselines, we considered two groups of ML approaches. First, we included Logistic Regression, $K$-Nearest Neighbor, and Naïve Bayes to assess how well HIXPred performs compared to intrinsically interpretable ML approaches. We added to this group the configuration of HIXPred, which only comprised the static module, as this module was intrinsically interpretable. Second, we also compared HIXPred against the ML classifiers Random Forest, AdaBoost, and Gradient Boosting to assess how well it performs compared to black-box ML approaches. We added to this group the configuration of HIXPred, which only comprised the sequential module, as this module used a Bi-LSTM and therefore acted as a black box. Further, for a fair comparison, we performed a grid search for each baseline. Finally, more details to the experimental setup, such as optimized hyperparameters, can be found in the shared online repository.\footnote{https://github.com/fau-is/hixpred}

5.3 Interviews with Medical Experts

To assess the explanation quality, we aimed to interview the healthcare providers involved in the underlying use case. Thus, we contacted the authors of the event log, from whom we, unfortunately, did not receive an answer. Therefore, we conducted semi-structured interviews with two independent medical experts. Both medical experts are physicians in different hospitals in different fields (surgery and anesthesiology) with five to ten years of professional experience. The interviews were structured into four parts. First, we provided all relevant information on the use case. Second, we described the initial situation where healthcare providers only receive predictions as typically provided by DNNs. Third, we provided our artifact’s visual output and some additional descriptions of the explanation for the static attributes. Fourth, we did the same for the sequential attributes. Based on this, we asked questions in regard to usefulness, applicability, and trust in the predictions.

5.4 Results

5.4.1 Predictive Performance

The predictive performance of HIXPred and the baselines are summarized in Table 2. The table shows the predictive performance average over ten executions) and the standard deviation in brackets behind the measure (sorted by highest $AUC_{ROC}$). The best overall results are marked in bold. The highest $AUC_{ROC}$ value was observed for HIXPred. Further, HIXPred outperforms the baselines in terms of precision and F1-score for both classes.

5.4.2 Interpretation and Explanation Quality

HIXPred creates interpretation as a bar plot with model coefficients for static features and explanation as swarm plot with SHAP values for sequential features. Regarding interpretation, Figure 4 (left) shows the coefficient values from the last layer of the model, indicating the effect of the static features on the respective pathway outcome prediction. A total of 11 static features had a positive effect on the outcome prediction, whereas 12 had a negative effect. Feature $\text{DisfuncOrg}$ had the highest positive effect which means that a disfunction in an organ has been detected. Another example is $\text{Hypotensie}$ (hypotension). It had the second-highest positive effect because low blood pressure is a typical indicator of the initial stage of sepsis (O’Brien Jr et al., 2007). Therefore, both are decisive reasons for admitting patients to ICU.
Figure 4 (right) shows the SHAP values for six of the 16 sequential features, indicating the effect of these features on the prediction.\textsuperscript{6} The values of Leucocytes have a negative effect on predicting “admission to ICU”. For CRP, most values negatively affect the prediction. They are low and, therefore, do not indicate inflammation. LacticAcid values are high and affect the prediction because an increased lactic acidosis is a typical indicator for a sepsis or sepsis shock (Suterstrong and Walley, 2016). ER Triage generally has a low effect on the prediction. Infusions of IV Liquid and IV Antibiotics have mostly positive effects.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
ML Approach & Class & Support & Precision & Recall & F1-score & AUC\textsubscript{ROC} \\
\hline
HIXPred \small{(Seq. and static module)} & Positive & 214 & 0.617 (±0.057) & 0.173 (±0.039) & 0.269 (±0.053) & 0.750 (±0.019) \\
& Negative & 3,338 & 0.949 (±0.002) & 0.993 (±0.002) & 0.971 (±0.001) & \\
\hline
Interpretable Models & & & & & & \\
Logistic Regression & Positive & 214 & 0.382 (±0.000) & 0.136 (±0.000) & 0.200 (±0.000) & 0.713 (±0.000) \\
& Negative & 3,338 & 0.947 (±0.000) & 0.986 (±0.000) & 0.966 (±0.000) & \\
Static module of DNN architecture & Positive & 214 & 0.413 (±0.059) & 0.183 (±0.032) & 0.252 (±0.031) & 0.691 (±0.021) \\
& Negative & 3,338 & 0.949 (±0.002) & 0.983 (±0.005) & 0.966 (±0.002) & \\
\textit{K}-Nearest Neighbor & Positive & 214 & 0.152 (±0.000) & 0.075 (±0.000) & 0.100 (±0.000) & 0.571 (±0.000) \\
& Negative & 3,338 & 0.943 (±0.000) & 0.973 (±0.000) & 0.958 (±0.000) & \\
Naive Bayes & Positive & 214 & 0.000 (±0.000) & 0.000 (±0.000) & 0.000 (±0.000) & 0.522 (±0.000) \\
& Negative & 3,338 & 0.934 (±0.000) & 0.992 (±0.000) & 0.965 (±0.000) & \\
\hline
Black-Box Models & & & & & & \\
AdaBoost & Positive & 214 & 0.519 (±0.000) & 0.131 (±0.000) & 0.209 (±0.000) & 0.749 (±0.000) \\
& Negative & 3,338 & 0.947 (±0.000) & 0.992 (±0.000) & 0.969 (±0.000) & \\
Gradient Boosting & Positive & 214 & 0.226 (±0.001) & 0.187 (±0.000) & 0.204 (±0.000) & 0.734 (±0.001) \\
& Negative & 3,338 & 0.948 (±0.000) & 0.959 (±0.000) & 0.954 (±0.000) & \\
Random Forest & Positive & 214 & 0.000 (±0.000) & 0.000 (±0.000) & 0.000 (±0.000) & 0.730 (±0.019) \\
& Negative & 3,338 & 0.940 (±0.000) & 0.999 (±0.000) & 0.969 (±0.000) & \\
Sequential module of DNN architecture & Positive & 214 & 0.000 (±0.000) & 0.000 (±0.000) & 0.000 (±0.000) & 0.638 (±0.015) \\
& Negative & 3,338 & 0.940 (±0.000) & 0.999 (±0.000) & 0.969 (±0.000) & \\
\hline
\end{tabular}
\caption{Comparison of predictive performance.}
\end{table}

\textsuperscript{6} In this paper, we present SHAP values for six of the 16 sequential features because of limited space, and for 1,000 prefixes, to keep the figure clear. As the first three features Leucocytes, CRP, and LacticAcid are continuous, the points in the SHAP plot can be colored differently. By contrast, the last three features ER Triage, IV Liquid, and IV Antibiotics differ only between zero and one. Because we filter out SHAP values where a data feature has a value of zero, the points in the SHAP plot are shown in one color. Additionally, we only show SHAP values between -0.100 and -0.100 to present the general effect of these features on the prediction.
5.4.3 Interviews

Both interviewees (I1 and I2) find the prediction target valid and the prediction itself useful for practical applications and see multiple benefits in practice. “A prediction regarding normal care unit or ICU would be helpful, as further therapeutic steps can be initiated at an early stage and bed capacity can be calculated at an early stage, and in case of possible overcrowding, other clinics/wards can be contacted in time.” (I1) Besides operational benefits for the healthcare provider, patients could also receive suitable treatment earlier, because such predictions “could lead to an earlier admittance to the ICU.” (I2)

Further, both interviewees confirm that predictions without additional explanation hinder trust in those predictions: “For medical personnel, the question remains how certain this prediction is and on what particular data it is based.” (I1) Additionally, “it is important to know at least the basics of how [predictions] come together.” (I2) Furthermore, they experience a need to understand the reasons for such predictions: “I just always feel the need to understand how these predictive [...] parameters were determined in their importance.” (I1)

Therefore, in order to ensure application in practice, an explanation should be given to healthcare providers, as they trust the prediction more if they have an explanation for it: “In my opinion, these explanations are helpful to practitioners [...], and I would trust [the predictions] more.” (I2) Additionally, such predictions with explanations would support practitioners in their decision-making: “I doubt that physicians would rely completely on such a prediction. But I can imagine that such a decision can be made more quickly by quickly summarizing the most important parameters.” (I1)

Moreover, besides the usefulness of the explanation, interviewee 1 also mentioned the correctness of the explanation based on the given example. In Figure 4 (left), Hypotension is highly relevant for the prediction of Admission to ICU: “That does make sense in my mind and may be an indicator.” (I1)

In summary, both expert interviews confirm that our artifact indeed provides a sufficient degree of explainability to comprehend the prediction model’s internal decision logic. As such, we can state that HIXPred not only provides high predictive quality but also ensures transparent decision support that is required for healthcare environments.

6 Discussion

We focused on patient pathways to demonstrate the benefits of HIXPred in cases with realistic properties in critical settings. Particularly in the context of healthcare, interpretability and explainability are crucial to allow for practical applications. However, our proposal is not limited to the medical domain but can also be used to predict process-related outcomes in other domains, including static and sequential data, as is often the case with patient event log data in many business environments. Furthermore, we decided to model the static patient data through a linear part in the neural network and the sequence of activities through a Bi-LSTM. HIXPred can easily be tailored to other needs. For instance, if the feature representing an activity can vary greatly in its value (not only binary but with a real value), it may be interesting to model the activity through the linear part. This would ensure that during the application phase of HIXPred, previously unseen values can not lead to unexpected outputs of the model. However, most activities are represented by a binary feature — one if the activity occurs and zero otherwise — which rules out unexpected behavior caused by unseen values.

Contributions: Our artifact was designed in a way to overcome the limitations of previous work. Thus, we can summarize HIXPred’s contribution in terms of the four derived requirements, which are all fully met by our artifact’s design. First, HIXPred supports a completely automated, fact-based analysis of patient pathways using patients’ historical event data to avoid subjectivity and reduce manual efforts to a minimal level. Second, HIXPred distinguishes between an offline and an online component. As such, it provides proactive analysis capabilities during the execution of patient pathways after training a prediction model based on historical data. On this basis, healthcare providers can proactively determine critical pathway
Hybrid Interpretable and Explainable Predictions

outcomes and detect weaknesses in advance for better resource allocation. Third, HIXPred provides high predictive performance without relying on explicit process knowledge. This allows flexibility for the application in highly complex and dynamic healthcare environments. Forth, HIXPred handles the problem of non-comprehensible predictions by providing sufficiently transparent results, which we consider the most crucial contribution of our work.

**Implications:** HIXPred combines two streams of research. On the one hand, we agree with the one stream, pushing towards inherently interpretable models (e.g., Rudin, 2019), that explanations of complex models can never provide the same understanding as inherently interpretable models. Thus, explanations should only be applied carefully. On the other hand, we also find that more complex models, such as sequential neural networks, can be a powerful choice to naturally model the data and, thereby, increase the predictive performance. However, these models come with the drawback of not being interpretable and requiring explanations, such as with SHAP. Thus, HIXPred makes use of fully interpretable parts in a neural network where possible, as well as complex model elements where the data type requires it.

Further, our research has multiple implications for information systems in healthcare management. First, the performance of ML models in predicting future patient events in patient pathways can be improved by combining patients’ static features with sequential data. This is a great advantage because prediction tasks in the healthcare sector are usually dominated by linear and logistic regression models with underlying static features to ensure a high degree of transparency. Therefore, we show that specifically carved neural networks can include both interpretable parts that behave similarly to traditional linear models and explainable parts that allow for more complex relations between the model’s inputs and outputs. If an increase in predictive performance is desired, the latter part is necessary to cover more complex structures while ensuring a sufficiently transparent output with the described limitations of a post-hoc XAI approach. In this regard, our evaluation and demonstration showed that HIXPred provides explanations that are helpful for a better understanding of the model’s predictions.

**Limitations and Future Research:** Similar to other research, our work is not free of limitations. First, the application of HIXPred was limited to a single case in order to present comprehensive evaluation results. However, in the next iteration of this DSR project, we will extend the application of HIXPred to additional real-world settings, where we also plan to evaluate the visual output of HIXPred with additional medical experts in a broader user-centric evaluation study. Also, we plan to investigate different types of interpretation and explanation representations, e.g., in the form of text, that might enhance HIXPred’s adoption in practice. Second, the sample size of the patient event log was relatively small and, therefore, may lack representativeness. However, neural networks have shown to improve predictive performance when trained with more data. Similarly, the data in the patient event log were collected over years and, therefore, may be subject to underlying concept drifts. Nevertheless, HIXPred has shown sufficient robustness to produce favorable results over a considerable period of time. A last limitation concerns the restricted generalizability of our results that goes beyond the developed artifact at hand. Thus, our ultimate goal is to abstract from the concrete implementation of the algorithmic artifact by deriving prescriptive design knowledge in the form of more generalizable design principles.

7 Conclusion

In this paper, we proposed HIXPred, a novel artifact that predicts the outcomes of patient pathways during run-time with high predictive performance and provides meaningful interpretations and explanations of these predictions for healthcare providers. To this end, we followed a DSR approach. More specifically, we first derived requirements based on current limitations in related research areas. Then, we designed our artifact as informed by prior research on XAI in general and explainable PBPM in specific. By instantiating HIXPred and applying it to a real-life healthcare use case, we evaluated and demonstrated its utility. Further, we conducted interviews with medical experts. Our findings confirm high predictive performance while ensuring sufficient interpretability and explainability to provide comprehensible decision support.
References


Hybrid Interpretable and Explainable Predictions


