



Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture

Julian Theis , *Graduate Student Member, IEEE*, William L. Galanter, Andrew D. Boyd, and Houshang Darabi , *Senior Member, IEEE*

Abstract—Diabetes intensive care unit (ICU) patients are at increased risk of complications leading to in-hospital mortality. Assessing the likelihood of death is a challenging and time-consuming task due to a large number of influencing factors. Healthcare providers are interested in the detection of ICU patients at higher risk, such that risk factors can possibly be mitigated. While such severity scoring methods exist, they are commonly based on a snapshot of the health conditions of a patient during the ICU stay and do not specifically consider a patient's prior medical history. In this paper, a process mining/deep learning architecture is proposed to improve established severity scoring methods by incorporating the medical history of diabetes patients. First, health records of past hospital encounters are converted to event logs suitable for process mining. The event logs are then used to discover a process model that describes the past hospital encounters of patients. An adaptation of Decay Replay Mining is proposed to combine medical and demographic information with established severity scores to predict the in-hospital mortality of diabetes ICU patients. Significant performance improvements are demonstrated compared to established risk severity scoring methods and machine learning approaches using the Medical Information Mart for Intensive Care III dataset.

Index Terms—Process mining, deep learning, in-hospital mortality, risk assessment, diabetes, intensive care.

I. INTRODUCTION

WITH 26.9 million diagnosed patients in the United States [1] and by accounting for 45% of ICU patients above the age of 65 [2], diabetes mellitus (DM) is a widespread illness that requires unique attention. There are two main types of DM. Type I is controlled with diet and insulin. Type II, the

more common, is controlled with diet, exercise, and a multitude of medicines. However, even a small infection can lead to severe patient outcomes and difficulties to control the disease. Hence, DM is rarely a standalone reason for severe patient outcomes but a factor that increases the likelihood of potentially life-threatening outcomes. In the aftermath, hospitalized DM patients require significantly more healthcare resources compared to other chronic disease populations [2]. One way to allocate healthcare resources more efficiently and to lower the rate of mortality is a precise in-hospital mortality risk assessment of diabetes ICU patients.

The calculation of risk and mortality scores in hospital settings has a long-standing history and is well-studied in the literature. The first methods were developed in the 1980 s and are used in healthcare facilities for decades. Commonly, such methods target the general patient population and include comorbidity assessment methods, such as the Elixhauser comorbidity score [3] and the Charlson comorbidity index (CCI) [4], and specific risk assessments like the probability for organ failure. Risk and mortality scores are calculated mostly based on patient information of the current admission and provide a snapshot of the patient's condition. With the ongoing adoption of Electronic Health Records (EHRs), healthcare facilities are building empiric patient data repositories. This enables the application of data mining and machine learning approaches. However, recent methods often neglect EHR data of patient's past hospital encounters when assessing their risk, or require substantial financial investment to be applicable in real-world.

A patient's health records can be understood as a sequence of observations. Such observations may include performed services, diagnoses, or lab measurements, and are also known as *careflows*. Process mining is a comparatively young research discipline that aims to extract knowledge from such sequences. Applications of process mining can be found across many industries, mainly to analyze and optimize applied processes, such as in business process management [5], automation [6], [7], manufacturing [8], [9], and recently in healthcare [10], [11]. Healthcare organizations increasingly acknowledge process mining and the use of empirical data to improve processes [12]. However, there is a lack of studies that include the patient's past hospital encounters using process mining to predict outcomes.

Manuscript received December 28, 2020; revised May 4, 2021; accepted June 19, 2021. Date of publication June 28, 2021; date of current version January 5, 2022. (Corresponding author: Houshang Darabi.)

Julian Theis and Houshang Darabi are with the Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: jtheis3@uic.edu; hdarabi@uic.edu).

William L. Galanter is with the Departments of Medicine, Pharmacy Practice, Pharmacy Systems, Outcomes and Policy, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: billg@uic.edu).

Andrew D. Boyd is with the Department of Biomedical, and Health Information Sciences, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: Boyda@uic.edu).

Digital Object Identifier 10.1109/JBHI.2021.3092969

In this paper, a novel process mining/deep learning architecture is proposed to enhance established risk calculation methods of in-hospital mortality of diabetes ICU patients by incorporating patient careflows from earlier hospital admissions. More specifically, the paper demonstrates a strategy to convert electronic health records to a careflow format suitable for process mining and how this information can be used to predict the patient outcome. Next to the past hospital encounters, the proposed approach leverages demographic information, diagnoses and procedures, diabetes-related health measurements, and existing risk scores that are calculated after 24 hours of admission.

The contributions of this paper are two-fold. First, a process mining/deep learning architecture is proposed to transform careflows embedded in EHRs into event logs suitable for process mining. Second, the proposed architecture is successfully demonstrated to improve the in-hospital mortality prediction of diabetes ICU patients by enhancing established risk scoring methods. The manuscript accentuates the non-negligible importance of modeling past patient careflows for outcome prediction and highlights process mining as a prospective set of tools for future research directions.

The paper is structured as follows. The related work is discussed in Section II. The fundamental preliminaries are provided in Section III followed by the proposed methodology in Section IV. Experimental evaluation and comprehensive result interpretations are reported in Section V. Section VI concludes the paper and provides future research directions.

II. RELATED WORK

The methods that are of interest for this paper can be separated into three categories: *Severity Scores*, *Data Mining and Machine Learning*, and *Process Mining*. Each category is introduced separately in this section.

A. Severity Scores

A common approach to assess the likelihood of in-hospital mortality of ICU patients is based on comorbidities. The *Elixhauser Score* [3] and CCI [4] are well-adopted in practice and calculated predominantly on diagnosis codes of patients [13]. The *Elixhauser Score* is the summation of points that are assigned if diagnoses belong to certain categories. In comparison, CCI weights more serious and more advanced conditions with more points. The Elixhauser and CCI scores are mainly used to assess the comorbidities of a patient and predict long-term mortality but are also used to predict in-hospital mortality [14]–[17]. Since DM requires specialized attention compared to the general patient population, a DM specific severity measure called *Diabetic Complications Severity Index (DCSI)* has been developed [18], [19]. The calculation is similar to the Elixhauser score. To the best of the authors' knowledge, no study investigated if DCSI predicts in-hospital mortality.

One of the most widely used algorithms to predict the mortality of patients in the ICU is the *Acute Physiology and Chronic Health Evaluation (APACHE) II* algorithm [20]. This severity score requires 15 variables, including physiological measurements, and provides a snapshot of the recent conditions

of a patient. With the increasing availability of EHRs and the deployment of patient workflow management software, APACHE II can be calculated automatically. However, APACHE II does not leverage available historical patient information. The subsequent versions, APACHE III and IV, are refinements and improvements of APACHE II and require a significantly larger number of variables [21]. The increase in necessary data led to the continued use of the APACHE II algorithm in practice [22].

The *Simplified Acute Physiology Score (SAPS) II* [23] reduces the complexity of APACHE-II. Similarly, SAPS-II is a measure of patient severity of illness and calculated based on demographic information, vitals, and lab measurements. SAPS-II has a sigmoidal relationship between mortality and its score and is therefore used to predict in-hospital mortality of ICU patients. The subsequent release, SAPS-III, uses 17 variables, including physiological and disease manifestation variables [24].

With the *Oxford Acute Severity of Illness Score (OASIS)*, a scoring has been developed to predict outcomes like mortality while requiring as few variables as possible [25]. OASIS is calculated based on seven physiologic measurements, elective surgery, age, and length of stay prior to ICU.

A further severity score is the *Sepsis-Related Organ Failure Assessment (SOFA)* that provides a risk calculation for organ failure [26]. SOFA has not been designed to predict mortality specifically, however, there is a positive correlation between the score and in-hospital mortality [27], [28]. Many Ethics Committees of hospitals are using SOFA to help plan for limited resources of ICU beds and ventilators during COVID-19 surges [29]. This adds risk associated with using scores in a way that they were not designed for [30] and highlights the immediate practical importance of mortality prediction.

Like APACHE, SAPS, and OASIS, SOFA provides a snapshot of the recent conditions of a patient with minimal or no consideration of historically available data of patients. Additionally, these severity scores have been developed for a broad patient population and were not specifically designed for DM patients.

B. Data Mining and Machine Learning

With the increasing availability of EHRs and computational advancements, data mining and machine learning methods have been proposed to assess patient outcomes. Recently, Brajer *et al.* [31] prospectively evaluated XGBoost based models to predict the in-hospital mortality of adults at the time of admission. However, the model uses only data from the recent admission and does not consider patient trajectories from past hospital encounters. Additionally, the model aims to predict in-hospital mortality for all types of hospital admissions and does not focus on the special attention that is required for diabetes ICU patients. Similar works have been published [32]–[34]. However, these methods do not consider the patient's careflow history and do not focus on DM patients. Solely Rajkomar *et al.* [35] predict in-hospital mortality by training a deep learning model on patient's entire EHR histories. However, their approach requires to map all EHR data to *Fast Healthcare Interoperability Resources (FHIR)* format across healthcare sites. Moreover, the training of the model is computationally expensive and needs specialized

know-how to develop. This requires substantial investment and therefore reduces practicability and near-term applicability significantly.

Two studies have been focusing on Machine Learning approaches to predict the in-hospital mortality of diabetes ICU patients. Anand *et al.* [22] developed Logistic Regression (LR) and Random Forest (RF) models to assess the mortality risk in diabetes ICU patients. The proposed approach considers patient demographic data, diabetes-related lab measurements, diagnoses, and medications. Similarly, Convolutional Neural Networks (CNN) were applied to a similar set of variables [36]. None of the methods consider the careflows of past hospital encounters of patients, though prior admission information is known to be a factor for outcome prediction [37].

C. Process Mining in Healthcare

Process mining originates from the analysis of business processes by investigating recordings of performed actions. EHRs and patient workflow management systems enabled the adoption of process mining in the healthcare domain. In comparison to business processes, healthcare processes are highly dynamic, complex, increasingly multi-disciplinary, and ad-hoc [38] which results in numerous research challenges.

One of the first healthcare applications of process mining was published by Mans *et al.* [39] in 2008. This work demonstrates the applicability of process mining techniques to a gynecological oncology process. The process was analyzed from control flow, organization, and performance perspectives and detected a lack of structure in healthcare environments. Since then, process mining in healthcare has gained significant attention [11], [40]–[43] and is increasingly used to identify regular behavior, careflow variants, and exceptional medical cases in an a-posterior way. The outcomes of process mining in healthcare are used to discover different disease progressions, treatment variations, detect bottlenecks, and evaluate how well a healthcare site conforms with guidelines. Process mining is also increasingly utilized to analyze and detect careflows across multiple organizations [44]. Using empirical data to improve processes is increasingly acknowledged by healthcare organizations [12].

Process mining techniques have also been applied to analyze diabetes-related processes. The study of de Toledo *et al.* [45] discovered common DM disease patterns from diagnosis data that is generally recorded for administrative purposes. The authors demonstrate the applicability of process mining to detect meaningful patterns from a real-world dataset of DM patients. Conca *et al.* applied process mining on a DM dataset to investigate if differences in work coordination of caregivers exist and if differences lead to undesired patient outcomes [46]. A further study by Dagliati *et al.* analyzed DM patients in Italy to unveil frequent care patterns that describe the evolution of the disease [47].

While the literature recognizes that process mining can be used to obtain valuable insights into healthcare processes and careflows, no study exists that uses process mining to predict future patient outcomes. Such outcomes include in-hospital mortality of diabetes ICU patients.

III. PRELIMINARIES

A. Event Log

The subsequent definitions are based on [48]. An event $a \in \mathcal{A}$ is an instantaneous change of the state of a system. \mathcal{A} describes the set of all possible events. An event instance E is recorded for each occurrence of an event. E is a vector that consists of at least two elements: the label of the corresponding event and the timestamp of occurrence. E can contain further non-mandatory elements that describe for instance resources, people, or costs. Since events are instantaneous and the point probabilities in continuous probability distributions are zero, the timestamps of two events cannot be equal [48]. A trace $g \in \mathcal{G}$ is a finite and chronologically ordered sequence of event instances. \mathcal{C} describes the infinite set of all possible traces. An event log $\mathcal{L} \subset \mathcal{G}$ is a set of traces. $\mathcal{L}_{i,j}$ denotes the j th event instance of the i trace of an event log \mathcal{L} . The cardinality of an event log, $|\mathcal{L}|$, corresponds to the number of traces.

B. Petri Net

A Petri net is a mathematical model. It is a commonly used technique in process mining to represent processes. Events that are recorded in an event log can be represented by transitions in a Petri net. A Petri net model is at any given time in a certain state. Whenever a transition fires, i.e. an event is observed, the state of the Petri net changes from one to another until a final state is reached. The reader is referred to [49], [50] for formal Petri net introductions and definitions.

C. Process Mining

Process mining is a comparatively young research discipline that focuses on the analysis of processes using event logs [5]. Process mining methods can be traditionally classified into three categories: *Process Discovery*, *Conformance Checking*, and *Process Enhancement*.

Process Discovery is confronted with the automated extraction of process models, such as Petri nets, from a given event log \mathcal{L} . Ideally, such a process model reflects the behavior seen in the event log but also generalizes well. *Conformance Checking* describes the quality assessment of a process model with regards to an event log \mathcal{L} . Exemplary conformance checking metrics include the measurement of how well a process model allows for the behavior in an event log and metrics that measure if a process model allows for behavior beyond the one recorded in an event log. The category of *Process Enhancement* deals with the extension and improvement of process models using information from an event log.

The reader is referred to [5] for a formal process mining introduction. This paper is based on methods that are categorized as *Process Discovery* and *Conformance Checking*.

D. Decay Replay Mining

Decay Replay Mining (DREAM) [48] is a process mining based methodology that is used to predict future process events. The method extends the places of a discovered Petri net process

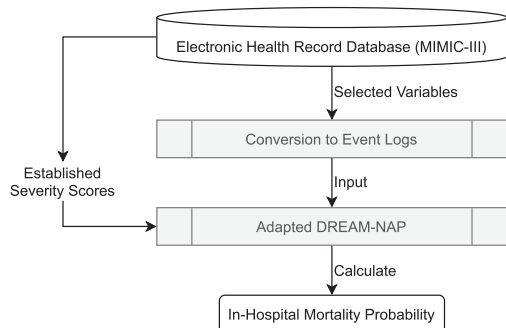


Fig. 1. High-level view of the proposed process mining/deep learning architecture to enhance the in-hospital mortality prediction of ICU diabetes patients.

model with time decay functions. These functions are parameterized using event timestamp information when replaying an event log on top of the process model. When replaying an event log on a Petri net that has been extended with time decay functions, one obtains a vector of time decay function values and a vector that represents the number of tokens that were created in each place next to the marking of the Petri net. The concatenation of marking, time decay function values, and token count is called *timed state sample* and represents the state of a process by incorporating time information. Additionally, a timed state sample is a lossfull embedding that can recover visited markings of replayed trace. The original publication [48] demonstrated the applicability of timed state samples to predict the next event given a partial trace and reports significant performance improvements.

E. Medical Information Mart for Intensive Care

The Medical Information Mart for Intensive Care III (MIMIC-III) is a large publicly available relational database that contains deidentified clinical data of patients that were admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts [51]. The database encompasses 38 597 adult patients and 49 785 hospital admissions from 2001 to 2012. The clinical data encompasses admission information, patient demographics, caregiver information, lab measurements, charted observations, full-text notes, diagnosis codes, and more. MIMIC-III has been recently evaluated and recommended for process mining purposes [52]. The database consists of multiple tables. Each table is defined as a relational variable T where $t_1, t_2, \dots, t_n \in T$ defines the column identifiers. A row of a table is defined by a vector R_T . The value r_t corresponds to the value of column t of row R .

IV. METHODOLOGY

This section focuses on the proposal of a process mining/deep learning architecture to enhance the in-hospital mortality prediction of diabetes ICU patients. The high-level overview is visualized in Fig. 1. The selection of variables is introduced in Section IV-A. In Section IV-B, an approach to convert EHRs to event logs is proposed. Finally, Section IV-C introduces

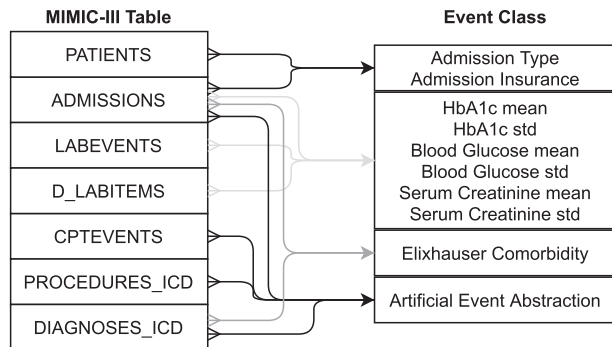


Fig. 2. This visualization provides an overview of the raw data sources required for each event class.

the methodology to predict the in-hospital mortality probability of DM ICU patients using an adapted DREAM-NAP approach [48]. The corresponding Python source code is publicly available on Github.¹

A. Variable Selection

The methodology has been developed using the MIMIC-III dataset. All DM patients were included that had at least one inpatient or outpatient hospital encounter registered in the database prior to their most recent ICU encounter.

The trace of each patient contains events of hospital admission types, admission timing, discharge times of past encounters, and insurance information. The possible insurance types are *Government*, *Self Pay*, *Medicare*, *Private*, and *Medicaid*. The admission types are *planned* and *unplanned*.

Furthermore, the hemoglobin A1C (HbA1c), serum creatinine, and blood glucose lab measurements are considered for each patient. These lab measurements are identifiable in MIMIC-III using the LOINC codes 4548-4, 2160-0, and 2345-7, respectively. The lab measurements are used regardless of inpatient or outpatient venue.

Additionally, the performed services and diagnoses are considered using CPT and ICD-9-CM codes. Codings are ordered in MIMIC-III by priority. A timestamp of a code is considered if present. Alternatively, the timestamp of a code is set to the timestamp of discharge.

Demographic information is used for each patient. This includes the age at the first registered encounter, sex, and race. Additionally, the SOFA, OASIS, APS-III, and SAPS-II score are calculated 24 hours after the admission for the index ICU encounter.

B. Health Records to Event Log Conversion

This section describes the conversion methodology to convert medical records from MIMIC-III to events that are suitable for process mining. In total, 10 different event classes are defined. Fig. 2 provides an overview how each table of MIMIC-III is used for the corresponding event classes. Events that are related to

¹[Online]. Available: <https://github.com/ProminentLab/PM-DL-Mortality-Prediction>

admission information is provided in blue. Green shows events that are based on lab measurements. Comorbidity events are visualized red. Artificial events resulting from diagnosis codes are highlighted in pink.

Initially, an empty trace g_p is created for each $p \in P$ where P is the set of relevant patients.

1) **Admission Events:** For each hospital admission of p , two events are created: one that reflects the type of admission with a timestamp equaling the recorded admission time, and one event that reflects the insurance of the patient with a timestamp that is delayed by 1 ms from the admission type. The delayed timestamp is calculated by the function $\sigma(t)$ where t is the initial timestamp. The delay is minimal. In this way, no information is altered and the delay is negligible given that the common time dimension in MIMIC-III is days.

For each hospital admission row $R_{T_{admissions}}$ of p , an event instance E is created such that e_{ts} equals the admission time of R . Moreover, an admission mapping is created that maps for each value of the set of values for $t_{admitttype}$ a value of the set $\{Planned, Unplanned\}$. The mapped value corresponds to the event name of E .

Additionally, a second event instance E is created for each admission row $R_{T_{admissions}}$ of p where e_{ts} equals the value of the function σ of the admission time of R . The event value equals the value in the insurance column. All events are added to g_p .

2) **Lab Events:** Lab measurements can be understood as time series data since each measurement might be performed multiple times during a specific hospital encounter or between hospital discharge and readmission. For each recorded time series and type of lab measurement, two events are created: a start event that corresponds to the observed mean value with a timestamp equaling to the timestamp of the first performed measurement, and an end event that represents the standard deviation (std) of the time series with a timestamp of the last performed measurement. If only one measurement was observed for a given time series, then the timestamp of the end event is set to 1 ms after the start event to maintain the order. Consequently, a lab event is calculated from one or many lab measurements. The corresponding events are added to g_p .

3) **Comorbidity Events:** Events are created to represent comorbidities. Therefore, the ICD-9-CM diagnosis codes of each hospital admission are used to calculate the Elixhauser comorbidity score. This score assigns points if certain ICD-9-CM codes are present representing a particular comorbidity category. If a point for such a category is assigned, a corresponding event is created such that the event name equals the comorbidity category. This leads to 30 possible different comorbidity events. The timestamp of comorbidity events corresponds to the discharge time of the corresponding hospital encounter. The comorbidity events of each patient are added to g_p . Since events in a trace occur sequentially, multiple comorbidity events with the same timestamps are delayed by multiples of 1 ms to maintain the order. The delayed timestamp is calculated by the function $\sigma(t)$ where t is the initial discharge timestamp.

4) **Artificial Events:** The time-ordered sequence of CPT and ICD-9-CM codes are used to create artificial events that carry essential, objective-dependent information. Since the number of unique ICD-9-CM and CPT codes is very large, a much

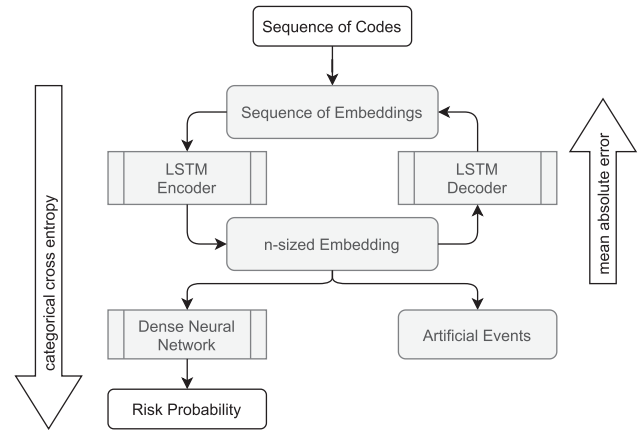


Fig. 3. This figure illustrates the flow to learn and extract artificial events from sequences of diagnosis and procedure codes.

smaller number of n artificial events need to be learned to transform the diagnosis and procedure codes into a space with a manageable dimension. Fig. 3 illustrates the methodology of learning artificial events from sequences of ICD-9-CM and CPT codes. Blue components represent neural network modules. These modules are connected from and to data objects. White data objects correspond to given inputs and desired outputs whereas gray objects represent latent data representations. The yellow arrows correspond to loss functions that are used to train the neural network objects that are visualized on the same vertical interception.

In a first step, the sequence of ICD-9-CM and CPT codes is transformed into a numerical space. Therefore, existing clinical embeddings of Choi *et al.* are leveraged [53]. The embeddings combine a large set of medical concepts including diagnoses, medications, procedures, and laboratory tests in one numerical space, and are learned using neural language modeling. These embeddings are derived from a private claim dataset of an health insurance company. This dataset contains ICD-9-CM codes, CPT procedure codes, medication, and lab measurement data of over four million patients longitudinally for 2-4 years per patient between the years of 2005 and 2013.

The embedded sequences are fed into a Long Short-Term Memory (LSTM) [54] that acts as an encoder with an output of n sigmoid neurons to learn an n -sized embedding of the sequence. The architecture of this encoder consists of two LSTM layers with each 100, and n LSTM cells and a dense output layer. Dropout is applied with a rate of 20% between all layers to prevent overfitting [55].

The resulting embedding is then fed as an input to a second LSTM to reconstruct the original embedded sequences using a mean absolute error loss. The architecture of the decoder is equal to the inversed architecture of the encoder. With this approach, an n -sized embedding of the sequences is learned that carries as much information as possible to reconstruct the input.

In parallel, the n -sized embedding is used to train a dense neural network to classify if the provided ICD-9-CM and CPT code sequence belongs to a patient that will either survive future hospital stays, is at risk of dying in a future hospital

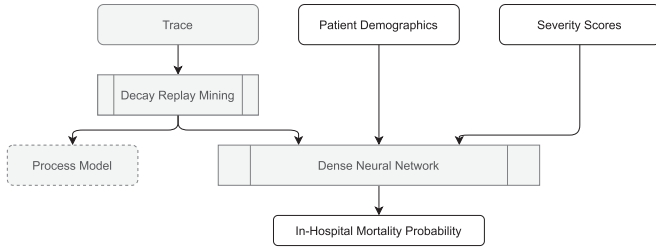


Fig. 4. Overview of the proposed approach to predict in-hospital mortality of diabetes ICU patients.

stays, or will die in the current hospital stay. This adds mortality information to the n -sized embedding when optimizing a categorical cross-entropy loss. At convergence, the n -sized embedding contains information to predict if a patient is at risk of in-hospital mortality while being able to reconstruct the original ICD-9-CM and CPT codes of importance. The dense neural network architecture consists of three layers with 35, 15, and 3 neurons, respectively. The first two layers leverage a Rectified Linear Unit activation [56] whereas the final layer uses a softmax activation to output the desired probabilities. Dropout of 20% is applied between all layers to prevent the model from overfitting.

Since the n -sized embedding consist of sigmoid outputs of the LSTM encoder, all n values are in the range of $[0,1]$. If a value at the index $i \in 0, 1, \dots, n-1$ is greater or equal than 0.5, an artificial event is created such that the name of the event equals index i . Hence, the existence of an event i corresponds to a value greater or equal than 0.5 at index i whereas the non-existence of an event i corresponds to a value smaller than 0.5 at index i of the n -sized embedding. The timestamps of artificial events equal to the timestamp of the sequence observation plus minimal offsets of 1 ms to maintain the order of the indices calculated by $\sigma(t)$. The resulting artificial events are added to the corresponding g_p .

C. In-Hospital Mortality Prediction

This section introduces a process mining based approach by leveraging the created event logs, patient demographics, and on-admission day severity scores to predict in-hospital mortality of diabetes ICU patients. An overview of the components is visualized in Fig. 4. Subsection IV-C1 introduces an extended DREAM approach [48]. Subsection IV-C2 focuses then on a dense neural network to predict the desired patient outcome.

1) *Adapted Decay Replay Mining*: DREAM demonstrated to be a promising process mining based methodology to predict the next event during the runtime of a process [48] by incorporating interarrival times of events using time decay functions. In this paper, two modifications to the original approach are proposed to predict in-hospital mortality events of diabetes ICU patients. The general idea is that the medical history of a patient is replayable on a process model. The timed state sample at the time of the most recent admission represents the state of the patient's comorbidities and diabetes-related health conditions on hospital presentation. As defined in [48], the timed state sample at time τ is defined as

$$S(\tau) = F(\tau) \oplus C(\tau) \oplus M(\tau) \oplus R(\tau) \quad (1)$$

which is the concatenation of the decay value vector $F(\tau)$, token count vector $C(\tau)$, marking $M(\tau)$, and resource counts $R(\tau)$ (if applicable) at time τ . In this paper, the timed state sample representation is modified to

$$S(\tau) = F_1(\tau) \oplus F_2(\tau) \oplus F_3(\tau) \oplus F_4(\tau) \oplus C(\tau) \oplus M(\tau) \quad (2)$$

where $F_1(\tau)$ corresponds to the original $F(\tau)$, and $F_{2-4}(\tau)$ correspond to novel time decay functions. When using several distinct types of time decay functions, one can highlight different time properties of the process when replaying the event log. This is assumed to provide a stronger learning signal for the prediction of events.

$F_1(\tau)$ and $F_2(\tau)$ contain time decay values of linear time decay functions, such as the one depicted in 3. However, the α parameter is initialized differently for the time decay functions in $F_2(\tau)$ compared to $F_1(\tau)$. Whereas the linear time decay functions of $F_1(\tau)$ are initialized using the mean reaction times of token, the functions of $F_2(\tau)$ are initialized based on the mean trace duration of \mathcal{L} .

$$f_1(\tau) = \beta - \alpha * \tau \quad (3)$$

$F_3(\tau)$ contains time decay values of exponential time decay functions that have been parametrized with the maximum trace duration. These functions are defined in 4. This type of time decay function provides a stronger signal on more recent place activations but vanishes faster compared to linear time decay functions. Time decay values have increased differentiability if they are occurring closer in time to each other.

$$f_3(\tau) = \beta * (1 - \alpha)^\tau \quad (4)$$

Finally, $F_4(\tau)$ is a vector that contains time decay values of logarithmic time decay functions. These functions are defined in 5 and are parametrized using the mean trace duration of \mathcal{L} .

$$f_4(\tau) = \log(\tau) / \log(\alpha) \quad (5)$$

Finally, an event count vector $A(g)$ is added that counts the occurrence of each event of a trace. Since the time decay function values, token counts, and markings are features that are obtained based on tokens entering and leaving places of the process model, the vector $A(g)$ contributes additional information of fired and visible transitions. This vector also counts all events of historical encounters and the index ICU encounter independent of time and sequence.

2) *Dense Neural Network*: The neural network required to predict the desired patient outcome has three input components that are different from the initially proposed neural network architecture in [48]: The timed states samples as an output from the adapted DREAM component described in Section IV-C1, demographic inputs and severity scores that are calculated on ICU admission day as a second input, and the event counts as a third input. An overview of the dense neural network architecture is provided in Fig. 5. The number in each layer represents the number of neurons. BN corresponds to a batch normalization layer and DO represents a dropout layer.

The timed state sample input layer takes $S(\tau)$ on admission as input. The second layer takes the age of a patient on its

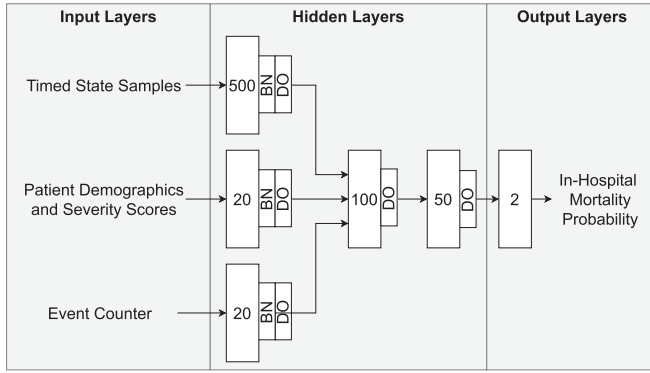


Fig. 5. Architecture of the dense neural network to predict in-hospital mortality of diabetes ICU patients.

first hospital encounter, gender, one-hot encoded ethnicity, and the severity scores (OASIS, SOFA, SAPS-II, and APS-III) that are calculated on ICU admission, concatenated into a single vector into consideration. The third layer considers the time- and sequence independent event count of a patient trace including events of the index encounter. This includes procedures, diagnoses, and diabetes-related health measurement values that are not necessarily available in the MIMIC-III database on admission, but that are commonly known and/or assessable by clinicians within 24 hours of ICU admission.

Each input layer is fed to a separated and unique hidden layer branch before a hidden layer concatenates the branches and feeds it to two further layers. All hidden layers use a Rectified Linear Unit activation function which has demonstrated superior performance in literature. The branched first hidden layers have additionally a batch normalization layer for accelerated convergence and improved stability [57] each. Dropout layers with a rate of 40% for regularization [55] are used. The output layer consists of a softmax activation function to output the patient's probability of in-hospital mortality. The detailed architecture is visualized in Fig. 5. The introduced architecture performed best across multiple hyperparameter selection iterations. Further insights are provided in Section V-D.

V. EVALUATION

This section describes the experimental evaluation of the proposed approach. The used dataset is described in subsection V-A followed by an introduction of the setup in V-B. The results and comparisons to existing methods is highlighted in V-C. Ablation studies are described in subsection V-D. A discussion in subsection V-E concludes Section V.

A. Dataset

DM patients with at least two hospital encounters where the most recent one encompassed an ICU encounter, are considered for evaluation. All of these patients must have had a DM-related ICD-9-CM diagnosis (250.xx) prior to the most recent admission. This excludes gestational DM and drug-induced DM. Furthermore, all patients were verified to either have an HbA1c value greater or equal than 6.5% [58] prior to the most recent admission, be prescribed DM home medication, or have had a

TABLE I
COHORT DETAILS OF THE TRAIN AND TEST SPLITS

Description		Train	Test
# Patients		1,827	609
Survival Rate		78.6 %	78.0 %
Average age [years]		65.8	66.4
Male Ratio		54.8 %	55.2 %
Mean # admissions/patient [years]		2.9	2.9
Mean history length [years]		2.3	2.4
Mean Encounter Duration [days]		10.2	9.5
Std Encounter Duration [days]		11.3	9.2
# Hospital Admission Types	Elective	127	36
	Emergency	1670	567
	Urgent	30	6
# Insurance Types	Medicare	1310	435
	Private	334	117
	Medicaid	149	50
	Government	29	7
	Self Pay	5	0

DM history in their medical history full text notes. This verifies and confirms that the selected patients are truly diabetic. This lead to a total number of 2436 patients which were randomly split into a training and testing cohort using a 75/25 ratio. The patient statistics of each split are described in Table I.

Additionally, 74 patients are added to the train set that were previously diagnosed with DM (250.xx) and went to the ICU, but that could not be verified to have an HbA1c value greater or equal than 6.5% prior to the most recent encounter, be prescribed diabetes home medication, or have had a diabetes history in their medical history full text notes. These patients are called *conjectural samples* and can be used for model training purposes only.

Furthermore, the train set is randomly split using an 85/15 ratio to obtain a train and validation split. These splits are required to discover a process model and train a neural network, and to select the best model before evaluating on the test split.

B. Setup

Process discovery and training of the described methodology is performed using the train set including the *conjectural samples* to maximize the training sample size. The *conjectural samples* are similar to the target group since these patients have been diagnosed with at least one diabetes-related ICD-9-CM code. Preliminary experiments have shown that using the set of *conjectural samples* for training increased the performance of predicting the in-hospital mortality of verified diabetes ICU patients using the proposed methodology.

The LSTM-based neural network architecture to learn artificial events that is introduced in Section IV-B4 is trained with an embedding size of $n = 30$. Hence, up to $n = 30$ artificial events are extracted from a given sequence of ICD-9-CM and CPT codes. The LSTM-based neural network has been trained for 300 epochs using a batch size of 128. The best model is obtained based on the performance on the validation set.

The corresponding timed state samples, demographic inputs, and severity scores are obtained for the train, test, and validation split. The train and validation splits are shuffled such that the sample sizes remain unchanged. In this way, the train split

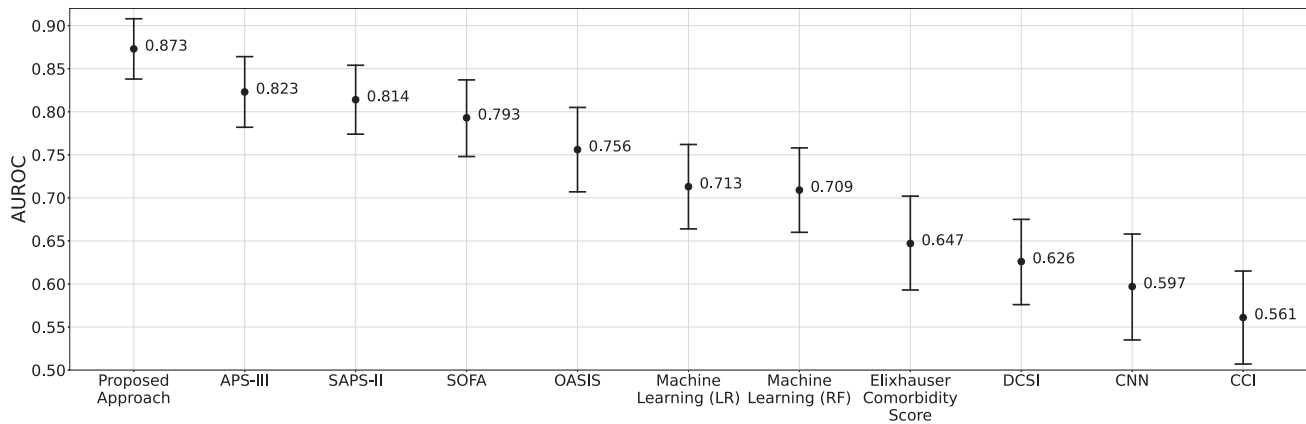


Fig. 6. This illustration provides a visualized overview of the obtained AUROC scores and their 95% CIs for each model.

contains samples that were unseen during process discovery and the validation set contains samples that have been used to obtain the process model. This increases the generalization of the neural network to predict in-hospital mortality. The neural network to predict the patient outcome is trained for 200 epochs using a batch size of 256. The best performing model is chosen based on validation loss and performance.

The test split is used to obtain the Area under the Receiver Operating Characteristic curve (AUROC). This score is an estimate of the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, and is a better classification estimate than other common classification performance metrics [59]. Moreover, the 95% confidence intervals (CIs) for the obtained AUROC scores are obtained using DeLong's method [60]. For evaluation and comparison purposes, the AUROC scores and CIs on the test set are obtained for the APS-III, SAPS-II, SOFA, OASIS, Elixhauser comorbidity, DCSI, and CCI scores. Moreover, LR and RF models are developed as described in [22]. A CNN-based model is trained as reported in [37] using the set of train patients as described in Section V-A.

C. Results

The proposed approach results in the highest observed AUROC score and the most narrow 95% CI. The severity score of APS-III demonstrates the second best AUROC score which is roughly 4.5% smaller than the one of the proposed approach with a CI that is more than 1% larger. SAPS-II shows also a satisfying performance in terms of AUROC. The SAPS-II AUROC is almost 1% smaller than APS-III, however, its CI is slightly narrower. The severity scores of SOFA, OASIS, DCSI, Elixhauser comorbidity score, and CCI show a decreasing predictive power in this order. At the same time, the CIs are increasing which implies an increasing uncertainty of each model when predicting the in-hospital mortality of diabetes ICU patients. Using the significant features that are described in [22] to build an LR and RF classifier result in lower AUROC scores with comparatively larger CIs than originally reported. It can be assumed that this is

due to the fact that the machine learning models were trained on patients that had at least two hospital encounters. Consequently, there are fewer patients available to train the model. Finally, the CNN model described in [36] performs with the lowest AUROC and the largest CI of all models. This score significantly differs from the originally reported score. The authors of [36] applied data transformations including oversampling prior to splitting patients into train and test cohorts. This leads to data leakage and causes consequently incorrect conclusions. In comparison, the evaluation scores reported in this manuscript are based on unseen patients contained in the test set and reflect the actual predictive performance of the model.

The CIs are visualized in Fig. 6. Since the CI of the proposed approach is non-overlapping with the CIs of OASIS, LR, RF, Elixhauser comorbidity score, DCSI, CNN, and CCI, the improvements are statistically significant. For APS-III, SAPS-II and SOFA, DeLong tests are performed with the null hypotheses that the ROC of the proposed approach is the same as the ROC of APS-III, SAPS-II, and SOFA, respectively. This statistical test leads to the corresponding and respective p-values of 0.0266, 0.0053, and 0.0006. The Holm-Bonferroni procedure [61] is used to deal with the familywise error rates. Since $0.0266 < 0.05$, $0.0053 < 0.025$, and $0.0006 < 0.0167$, it can be concluded that the proposed process mining/deep learning architecture performs significantly better than any of the compared methods at a level of $\alpha = 0.05$.

Furthermore, a derivation of the proposed model is trained that neglects the severity scores for in-hospital mortality prediction. This model results in an AUROC of 0.826 with a 95% CI of [0.783, 0.868]. This is numerically larger than the other models and statistically superior to all but the APS-III, SAPS-II and SOFA using Delong's test at $\alpha = 0.05$. Consequently, it can be concluded that a process mining approach, without leveraging existing severity scoring methods, results in an AUROC that is as good as APS-III, SAPS-II and SOFA for diabetes ICU patients. The corresponding ROC curves are visualized in Fig. 7. The proposed model improves especially in the low False Positive Rate area compared to other established methods.

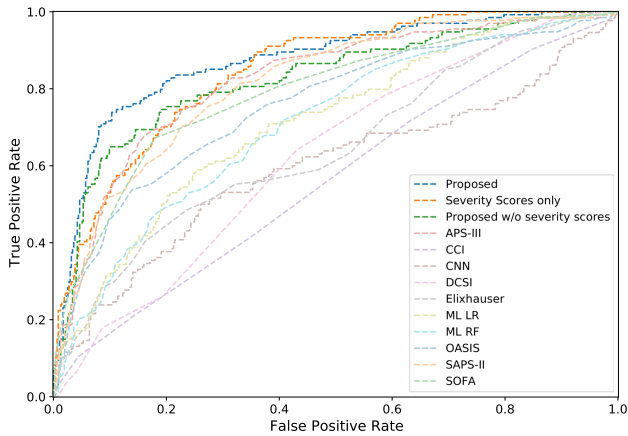


Fig. 7. ROC curves of the models.

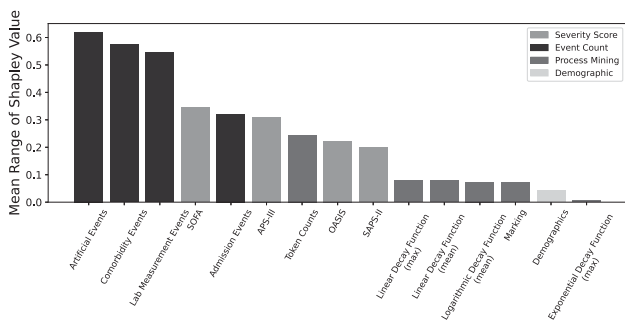


Fig. 8. Average 3 rd quartile range of Shapley values per feature type.

D. Ablation Study

This subsection describes multiple performed ablation studies, including an analysis of Shapley values of the neural network, an event type ablation on the event log, a layer-wise ablation study on the neural network, and an ablation study on the artificial event learning approach in the Sections V-D1, V-D2, V-D3, and V-D4, respectively.

1) *Shapley Value Analysis:* A Shapley value analysis is performed on every patient of the test set to investigate the impact of each patient’s input to the neural network output probability. Therefore, the Shapley Additive Explanation approach [62] is leveraged. A Shapley value describes the average contribution of a feature value to the prediction across different coalitions. The 3 rd quartile range of Shapley values provides a good estimate of a feature’s importance while neglecting outliers and accommodating for the sparsity of features with low impact due to the high dimensionality of the process model. Fig. 8 shows the obtained results.

The figure shows that the patient’s count of events has the largest impact on the prediction, followed by the severity scores. The token count vectors created by replaying the patient’s history on the process model have an impact similar to the severity scores. The average 3 rd quartile ranges of Shapley values for the time decay function values and markings of the timed state samples originating from the patient’s history show a range of around 10 percent. The demographic information and exponential time decay function seem to be less important on average. The figure confirms that patient history and the timing of events

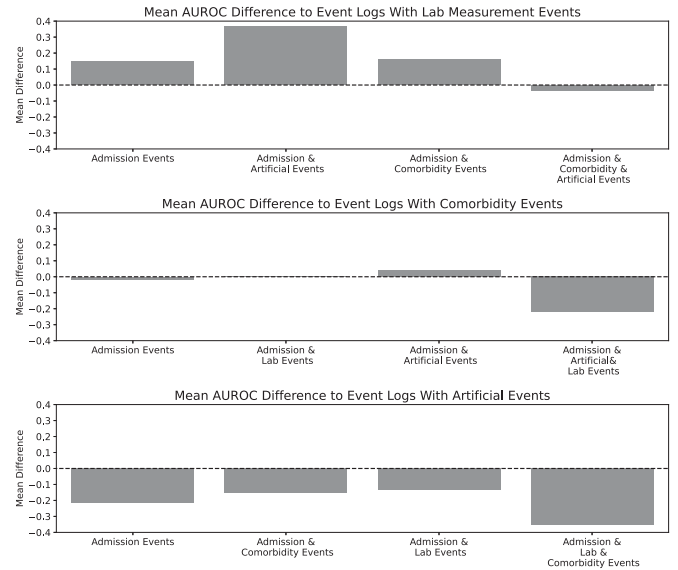


Fig. 9. Mean AUROC difference for including/excluding an event type over 10 runs. The reference value corresponds to the setup including the event type under investigation while the bars indicate the difference when excluding the event type under investigation.

modeled using process mining have an impact on the mortality probability prediction in the proposed setting.

2) *Event Type Ablation Study:* It has been demonstrated that the removal of severity scores and demographic information leads to a performance that is comparable to severity scores. The continued step-wise event removal from the event log is analyzed in this subsection. The impact of each event type is investigated by using all possible combinations of event logs that either include/not include lab measurement, comorbidity, and artificial events. Admission events are always required since those mark the extraction of timed state samples. This leads to a total of 8 unique event log setups that are trained 10 times each. Fig. 9 shows the mean AUROC differences compared to reference scores.

The results indicate a performance maximization when all event types are included. This justifies the incorporation of the proposed event definitions. Moreover, the plot highlights the presence of high-dimensional event interactions. The removal of an arbitrarily chosen event type from the proposed event log structure leads to lower predictive performance. At the same time, the inclusion of lab measurements improves the predictive performance only when both comorbidity and artificial events are present. In other scenarios, lab measurement events seem to be negligible. The learned artificial events have a high impact in every event log setting based on the bottom plot.

3) *Neural Network Layer Ablation Study:* The proposed neural network architecture is analyzed in two steps. First, an analysis is performed by adding further layers to each of the three inputs prior to the concatenation layer. Second, a step-wise decrease of layers after the concatenation layer is conducted. The results are visualized in Fig. 10.

Fig. 10 shows that a second layer per input prior to concatenation leads to a similar AUROC score over 10 runs, but with higher variance. With a third layer, the AUROC starts decreasing

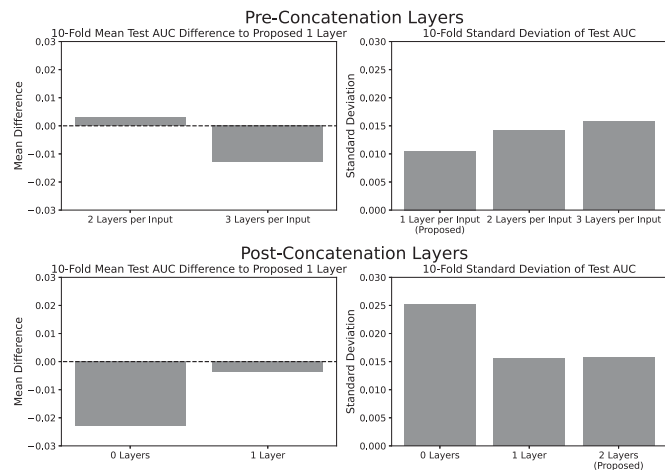


Fig. 10. Mean AUROC difference to proposed one-layer per input pre-concatenation and to proposed two-layer architecture post-concatenation including the corresponding standard deviations over 10 run.

with an even higher variability. This justifies one layer per input prior to concatenation of the proposed model. The figure shows also that the removal of all post concatenation layers (except for the softmax output) leads to an AUROC decrease with a large variance. Using one layer post concatenation results in a slightly worse AUROC performance with a variance compared to the proposed architecture.

The results indicate that the proposed architecture is locally optimized. Additionally, interactions across the different inputs are present since the architecture prior to the concatenation requires less layers compared to the one post concatenation.

4) *Artificial Embedding Analysis*: This subsection investigates the effectiveness of the proposed autoencoder architecture to learn artificial events, as described in Section IV-B4. The proposed architecture is compared to a reduced architecture with one hidden LSTM layer, and to an LSTM model where the embeddings are derived from the hidden state of the LSTM. Moreover, each architecture is trained with $n = 10, 20$, and 30 . The resulting nine architectures are trained and evaluated each ten times. The validation AUROC is used to interpret the performances. Fig. 11 visualizes the results.

The results show that the proposed architecture leads to constantly high validation AUROC scores independent of n . When increasing n , the comparison models increase in performance. Since all models with $n = 30$ lead to a similar validation AUROC with a similar variance, a deeper look at the resulting artificial events are required. The probability densities show that the proposed architecture leads to values that are closer to 0 or 1 with only few values in the mid-range around 0.5 compared to the simplified and the LSTM architecture. This shows that the proposed architecture learns the desired artificial events. When comparing the proposed architecture with $n = 30$ to $n = 20$, it can be observed that the value range is narrower and closer to 0.5. Hence, the usage of $n = 30$ is justified. An increase of $n = 40$ leads to a higher dimensionality of the process model that is less favorable.

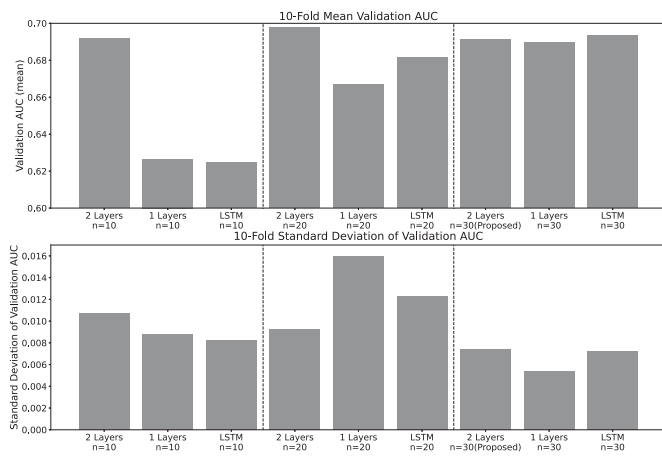


Fig. 11. 10-Fold validation set AUROC performance of three models with three embedding sizes.

E. Discussion

From the results, one can see that the severity scores that are calculated on ICU admission day - in particular SAPS-II, APS-III, SOFA, and OASIS - are good indicators for the in-hospital mortality of diabetes ICU patients. However, these scores are calculated based on a large set of present physiologic parameters with minimal or no consideration of prior patient histories. The proposed approach leverages those scores, also, and combines the information with information from past hospital encounters. This combination leads to an AUROC improvement of +5% from APS-III and +10% from OASIS. The Elixhauser comorbidity score, CCI, RF, LR, and CNN approaches can predict the in-hospital mortality of diabetes ICU patients with a weaker AUROC. However, these methods do not fully integrate past hospital encounters and leverage only a fraction of the available information.

One of the advantages of the proposed approach is the output of a process model that enables further analysis of patient careflows. A further advantage of using process mining over pure severity scores and traditional machine learning models is the discovery of a usually interpretable process model, as visualized in Fig. 4. This process model can be used for further traditional process mining analysis to gain insights into the patients at risk and their historical careflows.

A limitation of the proposed approach is that it requires that the patient have at least one prior admission and that the hospital has access to at least one of the last admissions. Hence, the approach is not suitable for patients which are previously only admitted to outside hospitals. Patients with DM who have never been admitted to a hospital before are likely less common in ICUs than those with prior admissions. In our dataset, roughly 40% of the DM patients in the ICU never had an encounter in the hospital before.

Therefore, the proposed approach addresses mainly large hospitals and hospital networks such that past hospital encounters of patients are available. Small hospitals that are geographically isolated may work well also as most admission(s) and other encounter data are captured. Small hospitals in urban areas

might lack the availability of data to build a process mining based model due to loss of encounter data. However, this limitation can be overcome with the integration of standards like FHIR to interchange medical health records between healthcare providers or through vendor initiated sharing networks like care everywhere² and Commonwell³.

VI. CONCLUSION

This paper has demonstrated one of the first process mining based approaches to model historical EHR data of diabetes ICU patients in combination with severity scores to predict in-hospital mortality. Specifically, an approach has been introduced that converts past medical records prior to the index hospital admission to event logs that are suitable for process mining. Then, a combination of existing risk scoring methods and Decay Replay Mining is used to predict the probability of mortality of a patient. In this way, established methodologies are combined with the advantages of incorporating historical information that provides an increased holistic view of the patients' conditions. The paper demonstrates significant performance improvements in predicting the in-hospital mortality of diabetes ICU patients that have a patient history in the hospital of the MIMIC-III database compared to established risk assessment scores and machine learning approaches.

The results underscore the importance of incorporating EHR information from the past into predictive systems. Furthermore, it demonstrates the suitability of process mining based methods to examine the impact of diabetes-related lab results, diagnoses, and procedures on outcomes in ICU patients.

However, the current methodology has also certain limitations. First, this approach addresses hospitals with a strong longitudinal patient record and may not be useful for ones with minimal longitudinal patient history. Yet, this limitation can be overcome if information exchange is occurring between the index and other medical centers. Second, the approach focuses on DM patients only and is not easily transferable to other chronic diseases. Each chronic disease may require its own process mining based model to maintain a satisfying objective-dependent predictive performance. Third, such an analysis using the MIMIC-III database can be difficult as patients procedures and diagnoses are ordered for billing purposes without the corresponding occurrence timestamp. Therefore, research on further hospital datasets is anticipated to strengthen the reported results and to validate the assumptions.

Future research is anticipated to be conducted in multiple directions. First, the proposed process mining/deep learning architecture enables multiple further healthcare applications that focus on various chronic disease patient groups and prediction tasks. This includes outcomes such as length of stay, ICU length of stay, or unexpected 30-day readmission on patients with chronic diseases like DM, kidney disease, or heart failure. Second, the proposed approach should be extended to analyze the

resulting process models of the DREAM approach for in-depth interpretability. Such an analysis can unveil further patterns that are predictors for severe patient outcomes in patient careflows.

REFERENCES

- [1] "National Diabetes Statist. Rep., 2020," Atlanta, GA: Centers for Disease Control and Prevention, US Dept. of Health and Human Services, 2020.
- [2] L. Fuchs *et al.*, "ICU admission characteristics and mortality rates among elderly and very elderly patients," *Intensive Care Med.*, vol. 38, no. 10, pp. 1654–1661, 2012.
- [3] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Med. Care*, pp. 8–27, 1998.
- [4] M. E. Charlson, P. Pompei, K. L. Ales, and C. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," *J. Chronic Dis.*, vol. 40, no. 5, pp. 373–383, 1987.
- [5] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 1st ed. Springer Publishing Company, Inc., 2011.
- [6] M. Haji and H. Darabi, "Petri net based supervisory control reconfiguration of project management systems," in *Proc. IEEE Int. Conf. Automat. Sci. Eng.*, 2007, pp. 460–465.
- [7] J. Theis and H. Darabi, "Behavioral petri net mining and automated analysis for human-computer interaction recommendations in multi-application environments," *Proc. ACM Human-Comput. Interact.*, vol. 3, no. EICS, pp. 1–16, 2019.
- [8] N. Wightkin, U. Buy, and H. Darabi, "Formal modeling of sequential function charts with time petri nets," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 2, pp. 455–464, May 2010.
- [9] J. Theis, I. Mokhtarian, and H. Darabi, "Process mining of programmable logic controllers: Input/output event logs," in *Proc. IEEE 15th Int. Conf. Automat. Sci. Eng.*, 2019, pp. 216–221.
- [10] H. Darabi, W. L. Galanter, J. Y.-Y. Lin, U. Buy, and R. Sampath, "Modeling and integration of hospital information systems with petri nets," in *Proc. IEEE/INFOSYS Int. Conf. Service Operations, Logistics Informat.*, 2009, pp. 190–195.
- [11] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *J. Biomed. Informat.*, vol. 61, pp. 224–236, 2016.
- [12] J. Lismont, A.-S. Janssens, I. Odnoletkova, S. vanden Broucke, F. Caron, and J. Vanthienen, "A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways," *Comput. Biol. Med.*, vol. 77, pp. 125–134, 2016.
- [13] H. Quan *et al.*, "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data," *Med. Care*, pp. 1130–1139, 2005.
- [14] C. van Walraven, P. C. Austin, A. Jennings, H. Quan, and A. J. Forster, "A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data," *Med. Care*, pp. 626–633, 2009.
- [15] M. E. Menendez, D. Ring, M. B. Harris, and T. D. Cha, "Predicting in-hospital mortality in elderly patients with cervical spine fractures: A comparison of the Charlson and Elixhauser comorbidity measures," *Spine*, vol. 40, no. 11, pp. 809–815, 2015.
- [16] M. E. Menendez and D. Ring, "A comparison of the Charlson and Elixhauser comorbidity measures to predict inpatient mortality after proximal humerus fracture," *J. Orthopaedic Trauma*, vol. 29, no. 11, pp. 488–493, 2015.
- [17] C.-Y. Kim, L. Sivasundaram, M. W. LaBelle, N. N. Trivedi, R. W. Liu, and R. J. Gillespie, "Predicting adverse events, length of stay, and discharge disposition following shoulder arthroplasty: A comparison of the Elixhauser comorbidity measure and Charlson comorbidity index," *J. Shoulder Elbow Surg.*, vol. 27, no. 10, pp. 1748–1755, 2018.
- [18] B. A. Young *et al.*, "Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization," *Amer. J. Managed Care*, vol. 14, no. 1, p. 15, 2008.
- [19] W. P. Glasheen, A. Renda, and Y. Dong, "Diabetes complications severity index (DCSI) - update and ICD-10 translation," *J. Diabetes Complications*, vol. 31, no. 6, pp. 1007–1013, 2017.
- [20] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: A severity of disease classification system," *Crit. Care Med.*, vol. 13, no. 10, pp. 818–829, 1985.
- [21] G. Niewiński, M. H. Starczewska, and A. Kański, "Prognostic scoring systems for mortality in intensive care units-the APACHE model," *Anaesthesiol. Intensive Ther.*, vol. 46, no. 1, pp. 46–49, 2014.

²[Online]. Available: <https://www.himss.org/resource-environmental-scan/care-everywhere>

³[Online]. Available: <https://www.commonwellalliance.org/how-to-participate/alliance-members/>

- [22] R. S. Anand *et al.*, "Predicting mortality in diabetic ICU patients using machine learning and severity indices," *AMIA Summits Transl. Sci. Proc.*, vol. 2018, p. 310, 2018.
- [23] J. R. Le Gall *et al.*, "Mortality prediction using SAPS II: An update for French intensive care units," *Crit. Care Med.*, vol. 9, no. 6, p. R 645, 2005.
- [24] R. P. Moreno *et al.*, "SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part II: Development of a prognostic model for hospital mortality at ICU admission," *Intensive Care Med.*, vol. 31, no. 10, pp. 1345–1355, 2005.
- [25] A. E. Johnson, A. A. Kramer, and G. D. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy," *Crit. Care Med.*, vol. 41, no. 7, pp. 1711–1718, 2013.
- [26] A. E. Jones, S. Trzeciak, and J. A. Kline, "The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation," *Crit. Care Med.*, vol. 37, no. 5, p. 1649, 2009.
- [27] J.-L. Vincent *et al.*, "Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study," *Crit. Care Med.*, vol. 26, no. 11, pp. 1793–1800, 1998.
- [28] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J.-L. Vincent, "Serial evaluation of the SOFA score to predict outcome in critically ill patients," *JAMA*, vol. 286, no. 14, pp. 1754–1758, 2001.
- [29] R. D. Truog, C. Mitchell, and G. Q. Daley, "The toughest triage - allocating ventilators in a pandemic," *New England J. Med.*, vol. 382, no. 21, pp. 1973–1975, 2020.
- [30] R. A. Raschke, S. Agarwal, P. Rangan, C. W. Heise, and S. C. Curry, "Discriminant accuracy of the SOFA score for determining the probable mortality of patients with COVID-19 pneumonia requiring mechanical ventilation," *JAMA*, vol. 325, no. 14, pp. 1469–1470, 2021.
- [31] N. Brajer *et al.*, "Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission," *JAMA Netw. Open*, vol. 3, no. 2, pp. e1920 733–e1920 733, 2020.
- [32] Y. P. Tabak, X. Sun, C. M. Nunez, and R. S. Johannes, "Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (ALaRMS)," *J. Amer. Med. Informat. Assoc.*, vol. 21, no. 3, pp. 455–463, 2014.
- [33] G. J. Escobar, M. N. Gardner, J. D. Greene, D. Draper, and P. Kipnis, "Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system," *Med. Care*, pp. 446–453, 2013.
- [34] C. T. Nakas, N. Schütz, M. Werners, and A. B. Leichte, "Accuracy and calibration of computational approaches for inpatient mortality predictive modeling," *PLoS One*, vol. 11, no. 7, 2016.
- [35] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, p. 18, 2018.
- [36] I. Wittler, X. Liu, and A. Dong, "Deep learning enabled predicting modeling of mortality of diabetes mellitus patients," in *Proc. Pract. Experience Adv. Res. Comput. Rise Machines (Learning)*, Ser. PEARC '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–6.
- [37] N. Schwartz, A. Sakhni, and N. Bisharat, "Predictive modeling of inpatient mortality in departments of internal medicine," *Intern. Emerg. Med.*, vol. 13, no. 2, pp. 205–211, 2018.
- [38] Á. Rebuge and D. R. Ferreira, "Business process analysis in healthcare environments: A methodology based on process mining," *Inf. Syst.*, vol. 37, no. 2, pp. 99–116, 2012.
- [39] R. S. Mans, M. Schonenberg, M. Song, W. M. van der Aalst, and P. J. Bakker, "Application of process mining in healthcare—A case study in a dutch hospital," in *Proc. Int. Joint Conf. Biomed. Eng. Syst. Technol.*, Springer, 2008, pp. 425–438.
- [40] E. Helm, A. M. Lin, D. Baumgartner, A. C. Lin, and J. Küng, "Adopting standard clinical descriptors for process mining case studies in healthcare," in *Proc. Int. Conf. Bus. Process Manage.*, Springer, 2019, pp. 608–619.
- [41] R. Williams, E. Rojas, N. Peek, and O. A. Johnson, "Process mining in primary care: A literature review," *Stud. Health Technol. Informat.*, vol. 247, pp. 376–380, 2018.
- [42] T. Erdoğan and A. Tarhan, "Process mining for healthcare process analytics," in *Proc. Joint Conf. Int. Workshop Softw. Meas. Int. Conf. Softw. Process Product Meas.*, 2016, pp. 125–130.
- [43] M. Ghasemi and D. Amyot, "Process mining in healthcare: A systematised literature review," *Int. J. Electron. Healthcare*, vol. 9, no. 1, pp. 60–88, 2016.
- [44] J. G. Garcia, C. Telleria-Orrriols, F. R. E. Romero, and E. Bernal-Delgado, "Construction of empirical care pathways process models from multiple real-world datasets," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 9, pp. 2671–2680, Feb. 2020.
- [45] P. de Toledo, C. Joppien, M. P. Sesmero, and P. Drews, "Mining disease courses across organizations: A methodology based on process mining of diagnosis events datasets," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 354–357.
- [46] T. Conca *et al.*, "Multidisciplinary collaboration in the treatment of patients with type 2 diabetes in primary care: Analysis using process mining," *J. Med. Internet Res.*, vol. 20, no. 4, p. e127, 2018.
- [47] A. Dagliati, V. Tibollo, G. Cogni, L. Chiovato, R. Bellazzi, and L. Sacchi, "Careflow mining techniques to explore type 2 diabetes evolution," *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 251–259, 2018.
- [48] J. Theis and H. Darabi, "Decay replay mining to predict next process events," *IEEE Access*, vol. 7, pp. 119 787–119 803, 2019.
- [49] W. van der Aalst, *Process Modeling and Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 55–88.
- [50] T. Murata, "Petri nets: Properties, analysis and applications," *Proc. IEEE Proc. IRE*, vol. 77, no. 4, pp. 541–580, Apr. 1989.
- [51] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, p. 160035, 2016.
- [52] A. P. Kurniati, E. Rojas, D. Hogg, G. Hall, and O. A. Johnson, "The assessment of data quality issues for process mining in healthcare using medical information mart for intensive care III, a freely available e-health record database," *Health Informat. J.*, vol. 25, no. 4, pp. 1878–1893, 2019.
- [53] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits Transl. Sci. Proc.*, vol. 2016, p. 41, 2016.
- [54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [56] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learning, Ser. ICML'10*. USA: Omnipress, 2010, pp. 807–814.
- [57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [58] S. E. Inzucchi, "Diagnosis of diabetes," *New England J. Med.*, vol. 367, no. 6, pp. 542–550, 2012.
- [59] C. X. Ling, J. Huang, and H. Zhang, "AUC: A better measure than accuracy in comparing learning algorithms," in *Proc. Conf. Can. Soc. Comput. Stud. Intell.* Springer, 2003, pp. 329–341.
- [60] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, pp. 837–845, 1988.
- [61] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, pp. 65–70, 1979.
- [62] S. M. Lundberg, and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. H. Bengio, R. Wallach Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.