Machine Learning (2023) 112:4409–4430
https://doi.org/10.1007/s10994-022-06300-x

Early anomaly detection in time series: a hierarchical approach for predicting critical health episodes

Vitor Cerqueira[1,2] · Luis Torgo[1,2,3] · Carlos Soares[2,3,4]

Received: 14 February 2020 / Revised: 29 October 2021 / Accepted: 30 December 2022 /
Published online: 27 February 2023

Abstract

The early detection of anomalous events in time series data is essential in many domains of application. In this paper we deal with critical health events, which represent a significant cause of mortality in intensive care units of hospitals. The timely prediction of these events is crucial for mitigating their consequences and improving healthcare. One of the most common approaches to tackle early anomaly detection problems is through standard classification methods. In this paper we propose a novel method that uses a layered learning architecture to address these tasks. One key contribution of our work is the idea of pre-conditional events, which denote arbitrary but computable relaxed versions of the event of interest. We leverage this idea to break the original problem into two hierarchical layers, which we hypothesize are easier to solve. The results suggest that the proposed approach leads to a better performance relative to state of the art approaches for critical health episode prediction.

Keywords Time series · Early anomaly detection · Healthcare · Layered learning

Editors: Petra Kralj Novak, Tomislav Šmuc.

✉ Vitor Cerqueira
vitor.cerqueira@dal.ca

Luis Torgo
ltorgo@dal.ca

Carlos Soares
csoares@fe.up.pt

[1] Dalhousie University, Halifax, Canada

[2] LIAAD-INESC TEC, Porto, Portugal

[3] University of Porto, Porto, Portugal

[4] Fraunhofer AICOS Portugal, Porto, Portugal

1 Introduction

1.1 Motivation for early anomaly detection in healthcare

Healthcare is one of the domains which has witnessed significant growth in the application of machine learning approaches (Bellazzi & Zupan, 2008). For instance, ICUs (Intensive Care Units) evolved considerably in recent years due to technological advances such as the widespread adoption of bio-sensors (Saeed et al., 2002). This lead to new opportunities for predictive modelling in clinical medicine. One of these opportunities is the early detection of critical health episodes (CHE), such as acute hypotensive episode (Ghosh et al., 2016) or tachycardia episode (Forkan et al., 2017) prediction problems. CHEs such as these represent a significant mortality risk factor in ICUs (Ghosh et al., 2016), and their timely anticipation is fundamental for improving healthcare.

CHE prediction can be regarded as a particular instance of early anomaly detection in time series data, also known as activity monitoring (Fawcett & Provost, 1999). The goal behind these problems is to issue accurate and timely alarms about interesting but rare events requiring action. In the case of CHE, a system should signal physicians about any impending health crisis.

One of the most common ways to address activity monitoring problems is to view them as conditional probability estimation problems (Fawcett & Provost, 1999; Tsur et al., 2018). Standard supervised learning classification methods can be used to tackle them. The idea is to approximate a function f that maps a set of input observations X to a binary variable y, which represents whether an anomaly occurs or not. In the context of CHE prediction, the predictor variables (X) summarise the recent physiological signals of a patient assigned to the ICU, while the target (y) represents whether or not there is an impending event in the near future.

1.2 Working hypothesis and approach

In many domains of application, the anomaly or event of interest is defined according to some rule derived from the data by professionals. In the case of healthcare, CHEs are often defined as events where the value of some physiological signal exceeds a pre-defined threshold for a prolonged period. Similar approaches for formalising anomalies can be found in predictive maintenance (Ribeiro et al., 2016), or wind power prediction (Ferreira et al., 2011). In these scenarios, we can also define pre-conditional events, which are arbitrary but computable relaxed versions of the event of interest. These pre-conditional events co-occur with the anomaly one is trying to model, but are more frequent and, in principle, a good indication for these. To be more precise, a pre-conditional event (i) represents a less extreme version of the anomalies we are trying to detect (main events); and (ii) co-occur with anomalies (i.e. there can not be an anomaly without a pre-conditional event). This concept is illustrated in the right-hand side of Fig. 3 as a Venn diagram for classes.

Our working hypothesis in this paper is that modelling these pre-conditional events can be advantageous to capture the actual events of interest. To achieve this, we adopt a layered learning method (Stone & Veloso, 2000). Layered learning denotes a hierarchical machine learning approach in which a predictive task is split into two or more layers (simpler predictive tasks) where the learning process within a layer affects the learning process of the next layer. This type of approach is also common in the hierarchical reinforcement learning literature (Dietterich, 2000a); for example, the options framework by Sutton (1998). Our contribution is its application to early anomaly detection problems.

Our approach exploits the idea that rare events of interest co-occur with pre-conditional events, which are considerably more frequent. Further, the same type of event of interest can be caused by distinct factors. For example, a particular type of CHE affecting two people may be caused by different diseases, which in turn may cause distinct dynamics in the time series of physiological signals. Therefore, initially modelling a relaxed version of the event of interest may lead to a simplification of the predictive task and a better performance when capturing the actual event of interest.

We apply the proposed approach to tackle the problem of predicting different types of CHE, including hypotension, hypertension, tachycardia, bradycardia, tachypena, bradypena, and hypoxia. Our results show that the layered learning model leads to a better average anticipation time for the same rate of false alarms when compared to different state of the art methods.

In short, the contributions of this paper are the following:

- a general hierarchical approach to the early detection of anomalies in time series data;
- the application of the proposed approach to several CHE problems, namely hypotension, hypertension, tachycardia, bradycardia, tachypena, bradypena, and hypoxia;
- a set of experiments validating the proposed approach, which includes a comparison with state of the art approaches.

This paper is structured as follows. In the next section, we start by formalising the problem of activity monitoring, both in general terms and using the case study of event prediction in ICUs. In sect. 3, we present the proposed layered learning approach to activity monitoring. We overview layered learning as proposed by Stone and Veloso (2000), and formalise our proposed adaptation for the early detection of CHE. In sect. 4, we carry out some experiments using the MIMIC II database (Saeed et al., 2002), and discuss these in Sect. 5. In sect. 6, we overview the related work, and finally conclude the paper in sect. 7.

We note that this paper is an extension of the work published in Cerqueira et al. (2019). The data is publicly available (Saeed et al., 2002). We also publish our code in an online repository[1].

2 Early anomaly detection

2.1 Formalization

We formalise the problem of early anomaly detection in time series in this section. We start by formalising the general problem and then the particular case of CHE prediction.

We follow (Weiss & Hirsh, 1998) to formalise the predictive task. Let $\mathcal{D} = \{D_1, \dots, D_{|\mathcal{D}|}\}$ denote a set of time series. In our case, $\mathcal{D}$ represents a set of patients being monitored at the ICU of an hospital. Each $D_i \in \mathcal{D}$ denotes a time series $D_i = \{d_{i,1}, d_{i,2}, d_{i,n_i}\}$, where n_i represents the number of observations for entity D_i, and each $d \in D_i$ represents information

[1] At https://github.com/vcerqueira/activity_monitoring_mimic.

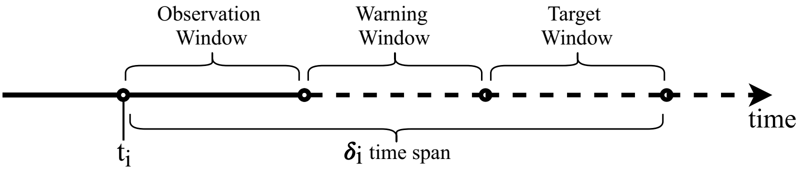


Fig. 1 Splitting a sub-sequence δ_i into observation window, warning window, and target window. The features X_i are computed during the observation window, while the outcome y_i is determined in the target window. In the timeline, the solid line describes the data available (past), while the dotted line represents the future

regarding D_i in the respective time step (e.g. a set of physiological signals captured from a patient in the ICU).

Each D_i can also be represented as a set of sub-sequences $D_i = \{\delta_1, \delta_2, \dots, \delta_i, \dots, \delta_{n'-1}, \delta_{n'}\}$, where δ_i represents the *i-th* sub-sequence. A sub-sequence is a tuple $\delta_i = (t_i, X_i, y_i)$, where t_i denotes the time stamp that marks the beginning of the sub-sequence, $X_i \in \mathbb{X}$ represents the input (predictor) variables, which summarise the recent past dynamics of the time series D_i; and $y_i \in \mathbb{Y}$ denotes the target variable, which is a binary value $(y_i \in \{0, 1\}, \forall\ i \in \{1, \dots, n'\})$ that represents whether or not there is an impending anomaly in the near future in the respective time series. How near in the future is typically a domain-dependent parameter. For each sub-sequence δ_i, we construct the feature-target pair (X_i,y_i) as follows.

As illustrated in Fig. 1, δ_i has three associated windows: (i) the target window (TW), which is used to determine the value of y_i; (ii) an observation window (OW), which is the period available for computing the values of X_i; and (iii) a warning window (WW), which is the lead time necessary for a prediction to be useful. An adequate WW enables a more efficient allocation of resources. Further, in the case of clinical medicine, physicians need some time after an alarm is launched to decide the most appropriate treatment.

The sizes of these windows depend on the domain of application and on the sampling frequency of the time series. In principle, the problem will be easier as the observation window is closer to the target window, that is, a smaller warning window is required. Weiss and Hirsh (1998) provide evidence for this property when predicting equipment failure, and Lee and Mark (2010) obtain similar results regarding hypotension prediction.

2.2 Event prediction in ICUs

In this work, we focus on a particular instance of early anomaly detection problems: CHE prediction in ICUs, particularly hypotension, hypertension, tachycardia, bradycardia, tachypena, bradypena, and hypoxia. For example, Ghosh et al. (2016) state that prolonged hypotension leads to critical health damage, from cellular dysfunction to severe injuries in multiple organs. Sustained tachycardia significantly increases the risk of stroke or cardiac arrest. Because CHEs are a relevant cause of mortality in ICUs, it is fundamental to anticipate them early in time so that physicians can prevent them or mitigate their consequences.

Patients assigned to the ICU are typically continuously monitored, with bio-sensors capturing several physiological signals, such as heart rate, or mean arterial blood pressure. This is illustrated in Fig. 2, where the data of a patient is depicted. A sub-sequence for

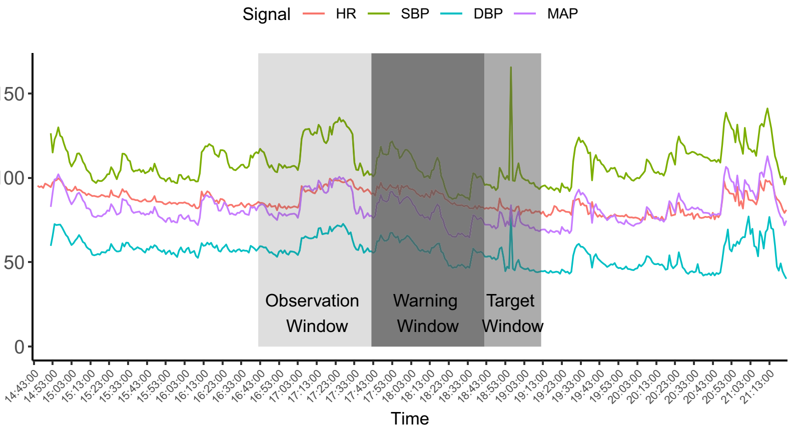


Fig. 2 The physiological signal of patients are monitored over time. Each sub-sequence, denoted by the shaded areas, is split in an observation window, a warning window, and a target window

Table 1 Critical health episodes addressed in this work. The threshold column defines the point from which the respective condition occurs

Event	Description	Threshold
Acute hypotension	Low mean arterial blood pressure	< 60 mmHg
Hypertension	High mean arterial blood pressure	> 105 mmHg
Tachycardia	High heart rate	> 100 beats p/ min (resting)
Bradycardia	Low hear rate	< 60 beats p/ min (resting)
Tachypena	High respiration rate	> 17 breaths p/ min
Bradypena	Low respiration rate	< 12 breaths p/min
Hypoxia	Low oxygen saturation	< 93%

CHE prediction is given as an example in the shaded area of the graphic. This area is split into three windows (observation, warning, target), as explained above.

Sensors capturing the physiological signals of a patient are typically collected with a high sampling frequency. In this work, we assume an instance arrives every minute for each ICU stay.

2.3 Critical health episodes

Table 1 describes each type of event we attempt to predict in this paper. We follow Forkan et al. (2017) closely to define these clinical conditions. Typically, these events occur when the respective physiological signal is below/above a pre-defined threshold for a prolonged period of time. For example, hypotensive episodes are defined as “a 30-minute window having at least 90% of its 6mean arterial blood pressure (MAP) values below 60 mmHg [millimetres of mercury]” (Tsur et al. 2018; Lee and Mark 2010). Essentially, this means that an hypotension episode occurs if the definition put forth in Table 1 is met for 90% of the instances in a given 30 minute period. We adopt this approach for all seven CHEs. In this context, the target variable value for a given CHE is computed as follows:

$$y_i = \begin{cases} 1, & \text{if a CHE happens in } TW_i, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, we consider that the i-th sub-sequence represents an anomaly if its target window represents a CHE (c.f. Fig. 2). Since CHEs are typically rare, the target vector y is dominated by the negative class (i.e. $y = 0$), where a patient shows a normal behaviour. For the target window of 30 minutes, we consider an observation window and a warning window of 60 minutes each.

2.4 Discriminating approaches to early anomaly detection

Naturally, one of the most common approaches to solving the problem defined previously is to view it as a conditional probability estimation problem and use standard supervised learning classification methods (Fawcett & Provost, 1999; Tsur et al., 2018). The idea is to build a model $f : \mathbb{X} \rightarrow \mathbb{Y}$, where $X \in \mathbb{X}$ and $y \in \mathbb{Y}$. This model can be used to predict the target values associated with unseen feature attributes. In other words, f is a discriminating model that explicitly distinguishes normal activity from anomalous activity.

Notwithstanding the widespread use of this approach, early anomaly detection problems often comprise complex target variables whose definition is derived from the data. In such cases, it is possible to decompose the target variable into partial and less complex concepts, which may be easier to model. In this context, our working hypothesis is that we can leverage a layered learning approach to model these partial concepts and obtain an overall better model for capturing the actual events of interest.

3 Layered learning for early anomaly detection

3.1 Layered learning

Layered learning is designed for predictive tasks whose mapping from inputs to outputs is complex. For example, Stone and Veloso (2000) apply this approach to robotic soccer. Particularly, one of the problems they face is the retrieval and passing of a ball. The authors split this task into three layers: (i) ball interception; (ii) pass evaluation; and (iii) pass selection. This process leads to a more effective decision-making system with a considerably higher success rate than a direct approach. This type of approach is also common in the reinforcement learning literature (Dietterich, 2000a). Specifically, hierarchical reinforcement learning consists in breaking a problem into a hierarchy of sub-tasks.

As Stone and Veloso (2000) describe, "the key defining characteristic of layered learning is that each layer directly affects the learning of the next". This effect can occur in several ways; for example, by affecting the set of training examples, or by providing features used for learning the original concept.

The general assumption behind decomposing a problem into hierarchical sub-tasks is that the problem addressed in each layer is simpler than the original one. We hypothesise that, when combining the models in each layer, this leads to a better overall approach for solving the task at hand.

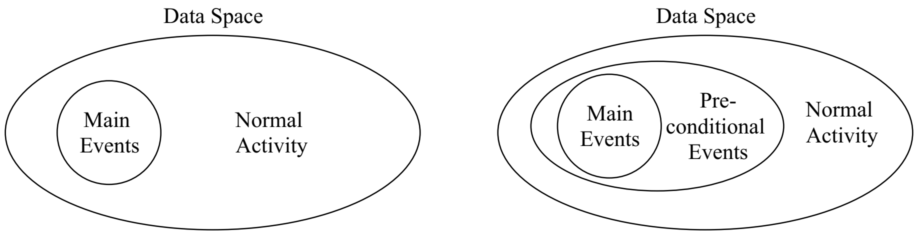


Fig. 3 Venn diagram for the classes in an activity monitoring problem. The main event represents a small part of the data space; pre-conditional events are more frequent and include the occurrence of the main events

3.2 Pre-conditional events

The definition of an anomalous event in time series data is in many cases determined according to some rule derived from the data. As an example from the healthcare domain presented in the previous section, a CHE is defined as a percentage of numeric values which are below some threshold within a time interval (c.f. sect. 2.3). This type of approach for defining anomalous events is also common in other domains. For example, in predictive maintenance (Ribeiro et al., 2016), numerical information from sensor readings is transformed into a class label which denotes whether or not an observation is anomalous. In wind ramp detection, a ramp event is a rare occurrence that denotes a large percentage change in wind power output in a short time interval (Ferreira et al., 2011).

Since these anomalous events are defined according to the value of an underlying variable, we can also define pre-conditional events: relaxed versions of the actual events of interest, but which are more frequent. A more precise definition can be given as follows. A pre-conditional event is an arbitrary but computable event that is expected to occur with the main event taking place simultaneously. If the main event occurs, the pre-conditional event must occur, but the latter can occur without the main event.

An example can be provided using the case study of hypotension prediction. According to sect. 2.3, we define the main event as “a 30-minute window having at least 90% of its mean arterial blood pressure (MAP) values below 60 mmHg”. A possible pre-conditional event for this scenario could be “a 30-minute window having at least 45% of its mean arterial blood pressure (MAP) values below 60 mmHg”. Another possibility is “a 30-minute window having at least 90% of its mean arterial blood pressure (MAP) values below 70 mmHg”.

In summary, pre-conditional events should have the following two characteristics:

- Pre-conditional events should have a higher relative frequency than the main events;
- Pre-conditional events always happen when the main events happen. The inverse is not a necessary condition. Another way to put it is that pre-conditional events are necessary, but not sufficient, events for the occurrence of the respective clinical condition.

3.3 Methodology

We can leverage the idea of pre-conditional events and use a layered learning strategy to tackle activity monitoring problems in time series data. Our idea is to decompose the main predictive task into two layers, each denoting a predictive sub-task. Pre-conditional events are modelled in the first layer, while the main events are modelled in the subsequent one.

The intuition behind this idea is given in Fig. 3. The figure presents two Venn diagrams for classes. Focusing on the left-hand side, the anomalies or main events (e.g. hypotension) represent a small part of the data space. This is one of the issues that makes them difficult to model. In the typical classification approach, main events are directly modelled with respect to the remaining data space (deemed normal activity).

Our idea is represented on the right-hand side. An initial pre-conditional concept is considered, which is more common than the main target concept, while also including it. The higher relative frequency of the pre-conditional events with respect to the main events helps to mitigate the problem of having an imbalanced distribution, which is the case in activity monitoring tasks. This phenomenon can compromise the performance of learning algorithms (He & Ma, 2013). In effect, we first model the pre-conditional events with respect to normal activity. These pre-conditional events are, in principle, easier to learn relative to the main concept because they are more frequent and thus the classification algorithms will not suffer so much from an imbalanced distribution. Afterwards, the main target events are modelled with respect to the pre-conditional events, which is also a less imbalanced distribution than the original on the left diagram.

In the remainder of this section, we will further formalise our approach using a generic notion of pre-conditional and main events. In the next section, we will apply this formalisation to CHE prediction problems.

3.3.1 Pre-conditional events sub-task

Let $\mathcal{S}$ denote a pre-conditional event. The target variable when modelling these events is defined as:

$$y_i^{\mathcal{S}} = \begin{cases} 1 & \textit{if } \mathcal{S} \text{ happens,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For this task, a sub-sequence $\delta_i^{\mathcal{S}}$ is a tuple $\delta_i^{\mathcal{S}} = (t_i, X_i, y_i^{\mathcal{S}})$. The difference to the original sub-sequence δ_i is the target variable, which replaces y with $y^{\mathcal{S}}$. Finally, the goal of this first predictive task is to build a function $f^{\mathcal{S}}$ that maps the input predictors X to the output $y^{\mathcal{S}}$.

3.3.2 Main events sub-task

Provided that we solve the pre-conditional events sub-task, in order to predict impending main events the remaining problem is to find out whether or not, when $\mathcal{S}$ happens, the main event also happens.

Let $\mathcal{F}$ be defined as the occurrence: "given $\mathcal{S}$, there is an impending main event in the target window of the current sub-sequence". Effectively, the target variable for this task is defined as follows:

$$\text{Given } y^{\mathcal{S}} = 1, y_i^{\mathcal{F}} = \begin{cases} 1 & \text{if a main event happens in } TW_i, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The target variable for this sub-task ($y^{\mathcal{F}}$) is formalised in equation 2. Given that the class of $y^{\mathcal{S}}$ is positive (which means that there is an impending pre-conditional event), the class of $y^{\mathcal{F}}$ is positive if a main event also happens in that same target window, or negative otherwise.

The goal of this second predictive task is to build a function $f^{\mathcal{F}}$, which maps X to $y^{\mathcal{F}}$. Formally, a sub-sequence $\delta_i^{\mathcal{F}}$ is represented by $\delta_i^{\mathcal{F}} = (t_i, X_i, y_i^{\mathcal{F}})$. In this scenario, however, the set of available sub-sequences D_i is considerably smaller than in the pre-conditional sub-task because only the sequences for which $y^{\mathcal{S}}$ equals 1 are accounted for. This means that $f^{\mathcal{F}}$ only learns with sub-sequences that at least lead to a pre-conditional event. Effectively, this aspect represents how the learning in the pre-conditional events sub-task affects the learning on the main events sub-task, i.e., by influencing the data examples used for training. In the main events sub-task, a predictive model is concerned with the distinction between pre-conditional events and main events. Essentially, it assumes that the distinction between normal activity and pre-conditional events is carried out by the previous layer. Given this independence, the training of the two layers can occur in parallel.

3.3.3 Predicting impending anomalies

To make predictions about impending events of interest we combine the models $f^{\mathcal{S}}$ with $f^{\mathcal{F}}$ with a function $g : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{Y}$.

$$g(X_i) = f^{\mathcal{S}}(X_i) \cdot f^{\mathcal{F}}(X_i) \tag{3}$$

Essentially, according to equation 3 the function g predicts that there is an impending main event in a given sub-sequence δ_i according to the multiplication of the outcome predicted by both $f^{\mathcal{S}}$ and $f^{\mathcal{F}}$.

Ideally, there are three possible outcomes:

- Both event $\mathcal{S}$ and event $\mathcal{F}$ happen, which means there is an impending main event: both $f^{\mathcal{S}}$ and $f^{\mathcal{F}}$ should return 1 so that $f^{\mathcal{S}} \cdot f^{\mathcal{F}} = 1$;
- Event $\mathcal{S}$ happens, but event $\mathcal{F}$ does not happen: $f^{\mathcal{S}} = 1$, but $f^{\mathcal{F}} = 0$, so $f^{\mathcal{S}} \cdot f^{\mathcal{F}} = 0$;
- Event $\mathcal{S}$ does not happen, and consequently, event $\mathcal{F}$ also does not happen: $f^{\mathcal{S}} \cdot f^{\mathcal{F}} = 0$.

3.4 Application of layered learning to CHE prediction

As mentioned before (c.f. sect. 2.3), a CHE is defined as a 30-min period where 90% of the respective physiological signal values are below or above the threshold. We propose to relax this threshold and define the pre-conditional event $\mathcal{S}$ as:

$\mathcal{S}^{\text{CHE}}$: “a 30-minute window having at least 45% of its physiological signal values below/above the pre-defined threshold".

We picked the value 45% arbitrarily (half the original percentage), and use it for all seven problems. Essentially, we attempted to make the pre-conditional events much more frequent relative to the main events. Nevertheless, this parameter can be optimised. Similarly to other hierarchical methodologies, for example in hierarchical reinforcement learning (Dietterich, 2000b), the definition of the sub-task is performed manually. In sect. 5.2, we will discuss this issue further.

4 Empirical experiments

The central research question addressed in this paper is the following:

How does the proposed layered learning strategy perform relative to other state of the art approaches for the early anomaly detection of critical health events?

4.1 Case study and predictive tasks

4.1.1 MIMIC II

In the experiments, we used the database Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC) II (Saeed et al., 2002), which is a benchmark for several predictive tasks in healthcare, including CHE prediction.

As inclusion criteria of patients and general database pre-processing steps, we follow (Lee & Mark, 2010) closely. For example, the sampling frequency of the physiological data of each patient in the database is minute by minute. Moreover, the following physiological signals are collected: heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial blood pressure (MAP), respiration rate (RR), and oxygen saturation (SPO2).

As described in sect. 2.2, the target window size is 30 minutes. For each target window, there is a 60-minute observation window and a 60-minute warning window. For a comprehensive read regarding the data compilation, we refer to the work by (Lee & Mark, 2010). Considering this setup, the number of patients is 2.643, leading to a data size of 10.067.577 sub-sequences.

4.1.2 Pre-processing and feature engineering

We consider HR, SBP, DBP, and MAP values between 10 and 200 (bpm for HR, mmHg for the remaining ones). Values outside of this range are eliminated as “unlikely outliers” (Lee & Mark, 2010). From the available signals (HR, SBP, DBP, MAP, RR, SPO2), we compute the values of cardiac output (CO) and pulse pressure (PP).

Regarding feature engineering, we follow previous work in the literature (Lee & Mark, 2010; Tsur et al., 2018). Using the observation window of each sub-sequence and of each physiological signal, the feature engineering process was carried out using statistical, cross-correlation, and wavelet functions. The statistical metrics include skewness, kurtosis, slope, median, minimum, maximum, variance, mean, standard deviation, and inter-quartile range. For each observation window, we also compute the cross-correlation of each pair of signals at lag 0. We also use the Daubechies wavelet transform (Percival & Walden, 2006) to perform a 5-level discrete wavelet decomposition and capture the relative energies in different spectral bands. Intuitively, medication data can play an important role. However, Lee and Mark (2010) reported no predictive advantage in using such information. In effect, we do not include this information in the predictive models.

4.1.3 Predictive tasks

We explore the research question put forth above in the seven problems described in sect. 2.3. These are also outlined in Table 2, which also includes the distribution of anomalous events and the distribution of the respective pre-conditional event.

Table 2 Distribution of each anomalous event and its respective pre-conditional event

Event	Event Dist. (%)	Pre-conditional Event Dist. (%)
Bradycardia	2.05	3.25
Bradypena	3.81	7.57
Hypertension	2.42	4.96
Hypotension	1.59	4.00
Hypoxia	5.76	9.95
Tachycardia	9.86	13.00
Tachypena	30.49	43.83

We apply the same data pre-processing steps and learning algorithms for all these predictive tasks.

4.2 Performance estimation

To estimate the predictive performance of each method, we used a 10-fold cross-validation, in which folds are split by patients. To be more precise, in each iteration of the cross-validation procedure, one fold of patients is used for validation, another fold of different patients is used for testing, and the remaining patients are used for training the predictive model. Therefore, all sub-sequences of a given patient are only used for either training, validation, or testing. The set of time series (patients) only comprises a temporal dependency within each patient, and we assume the data across patients to be independent. That is, the probability that a patient suffers a health crisis is independent of another patient also suffering a health crisis. In this context, the application of cross-validation in this setting is valid. Finally, the sub-sequences of the patients chosen for training are concatenated together to fit the predictive model. This model is tuned using the sub-sequences of patients chosen for validation and evaluated using the sub-sequences of patients chosen for testing.

4.2.1 Sub-sequences used for training

Given the sizes of OW, WW, and TW (60, 60, and 30 minutes, respectively), the duration of a sub-sequence is 150 minutes. Since the data is collected every minute, there is considerable overlap between consecutive sub-sequences. During run-time, a given model is used to predict whether there is an impending CHE in each sub-sequence. This approach emulates a realistic scenario, where a prediction is produced as more data is available regarding the current health state of a given patient.

Given the redundancy among consecutive sub-sequences, it is common to sample the sub-sequences for training a predictive model (Tsur et al., 2018). For example, Cao et al. (2008) compile sub-sequences for training according to whether a patient has experienced a CHE. For every patient that did, the latest 120 minutes of data before the onset of the respective CHE are used to create a training sub-sequence. If a patient did not experience a CHE, one or more sub-sequences are sampled at random. Lee and Mark (2010) collect multiple sub-sequences in a sliding window fashion, irrespective of whether a patient experienced a CHE. A sliding window with no overlap and of size TW is used to traverse each patient. That is, if a sub-sequence δ_i starts at time t_j then the next sub-sequence δ_{i+1} starts at time t_{j+30}. The authors show that this approach leads to better results relative to the approach taken by Cao et al. (2008). In both cases described above, the authors note that these approaches lead to an imbalanced data set. They recommend under-sampling the majority class to overcome this issue.

In this work, we follow the approach by Lee and Mark (2010). As recommended, we also apply a class balance procedure, which is described below in sect. 4.3.

4.2.2 The value of a prediction

The timely prediction of impending CHEs enables a more efficient allocation of ICU resources and a more prompt application of the appropriate treatment. In this context, for a prediction to be useful, it must occur before the onset of the respective CHE. We assume that, after the event starts, any prediction becomes obsolete. Further, predicting too early also leads to meaningless predictions due to the continuity of time. We follow the approach taken in the 10^{th} PhysioNet challenge (Moody & Lehman, 2009) regarding hypotension prediction, and extend it to all seven problems addressed in this work. A CHE is considered to be correctly anticipated if it starts within 60 minutes after an alarm is launched. We consider the value of an alarm to be binary, where its benefit is 1 if it is issued correctly, and 0 otherwise.

4.2.3 Learning algorithms

We tested different predictive models in the experiments, namely a random forest (Wright, 2015), a support vector machine (Karatzoglou et al., 2004), a deep feed-forward neural network (Abadi et al., 2016), an extreme gradient boosting (xgboost) model (Chen et al., 2015), and the lightgbm algorithm (Ke et al., 2017). We only show the results of the latter in these experiments, since it provides better performance than the remaining methods for the prediction of the CHEs addressed in this paper. Each learning algorithm is optimized on a validation set (c.f. 4.2) using a grid search.

4.3 State of the art methods

We compare the proposed layered learning approach (henceforth denoted as `LL`) with the following four methods.

4.3.1 Standard classification

We compare `LL` with a standard classification method (`CL`) that does not apply a layered learning approach and directly models the events of interest with respect to normal activity (c.f. Fig. 3). One of the working hypothesis for the application of the proposed layered learning approach is that it helps to mitigate the class imbalance problem. To further cope with this problem, we process the data used for training `CL` and `LL` using a re-sampling method. In the case of `LL`, this process was applied to both layers after performing the task decomposition. We applied SMOTE (Chawla et al., 2002) in both cases for all problems. We also tested other strategies in our experiments (e.g. random undersampling, random oversampling), but overall SMOTE performed better.

4.3.2 Isolation forest

An Isolation Forest (`IF`) (Liu et al., 2012) is a state of the art unsupervised model-based approach to anomaly detection. A typical method of this sort typically discards the anomalies within the training data and creates a model for normal activity. Observations that significantly deviate from the typical behaviour are considered outliers. We referred to these approaches as profiling methods (sect. 6.1). Instead of separating the normal activity, `IF` explicitly models the anomalies in an unsupervised manner using an ensemble of tree-structured models. The core idea behind a `IF` is that the paths resulting from partitioning the data are shorter for anomalous observations because the regions comprising these anomalies are separated quickly.

4.3.3 Ad-hoc methods

While there is an increasing number of machine learning applications in healthcare, many of the currently deployed systems still rely on simple *ad-hoc* rules to support the decision-making process of professionals. Taking hypotension prediction as an example, a simple rule is to trigger an alarm if the MAP of a patient drops below 60 mmHg in a given time step. A similar approach can be used for the remaining anomalous events, where an alarm is launched if the respective variable variable exceeds the threshold at any point in time. However simple, these ad-hoc rules often work well in practice. We use these rules as baselines in our experimental design and denote them as `AH`. The threshold for each CHE is described in Table 1.

4.4 Evaluation metrics

Approaches dealing with the prediction of critical health episodes typically evaluate predictive models using classical classification metrics, namely precision, recall, and F1 (Lee & Mark, 2010, 2010; Ghosh et al., 2016; Tsur et al., 2018). However, these metrics are unsuitable when dealing with high frequency time series because, as Fawcett and Provost (1999) state, they ignore the temporal order of observations and the value of timely decisions.

The goal behind early anomaly detection problems is not to classify each sub-sequence as positive or negative (Fawcett & Provost, 1999). Instead, the main goal is to detect, in a timely manner, when there is an impending anomalous event. In this context, we follow (Fawcett & Provost, 1999) regarding the evaluation process and apply the activity monitoring operating characteristic (AMOC) analysis. The rationale behind AMOC curves is similar to that of the well-known ROC curves, but tailored for time-dependent problems.

The AMOC curve used in this work differs from the standard ROC curve in two ways. First, it plots the expected number of false alarms, normalized by the unit of time, on the x-axis. We apply 60 minutes as the unit of time, which represents the size of the observation window. As an example, a false alarm rate (FAR) of 0.5 means that a false alarm is expected every 30 minutes. Second, the AMOC curve plots the expected value of the alarm on the y-axis instead of the true positive rate. As described in sect. 4.2.2, we quantify the value of an alarm according to anticipation time. Therefore, the expected value of the alarm represents the average anticipation time, which

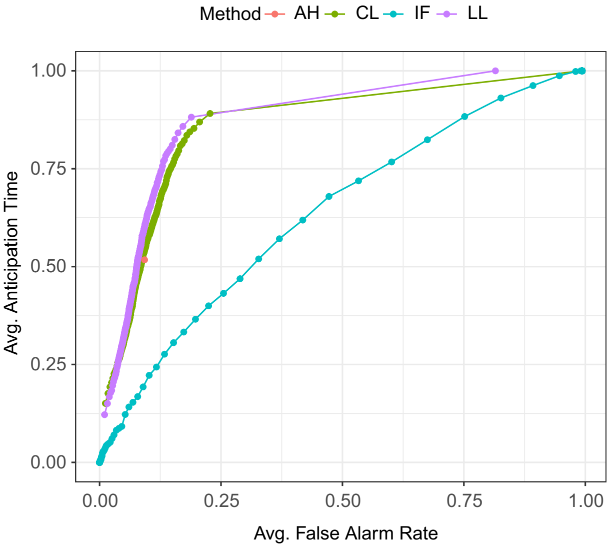


Fig. 4 AMOC curves for the bradycardia predictive task

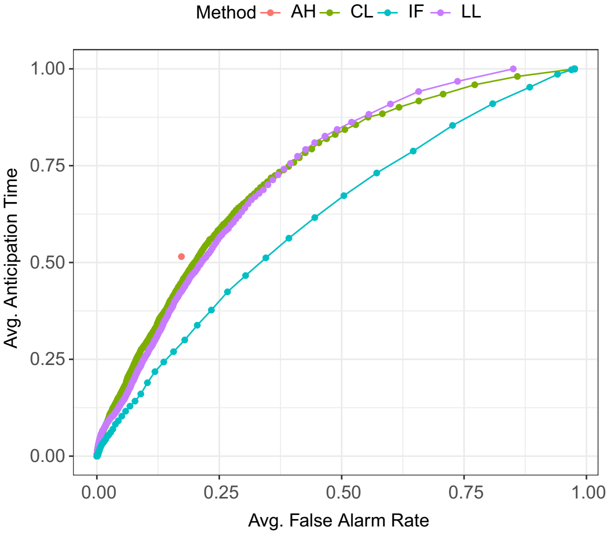


Fig. 5 AMOC curves for the bradypena predictive task

is also normalized by a period of 60 minutes. In effect, a score of 0.5 means that we expect to predict an event with 30 minutes in advance. We refer to the work by Fawcett and Provost (1999) for a comprehensive read on AMOC analysis.

As we mentioned before, we use a 10-fold cross-validation process to estimate predictive performance. The AMOC curve for each method is created by concatenating the results across all folds. Therefore, each curve presented below reflects the predictive performance of each method across all patients.

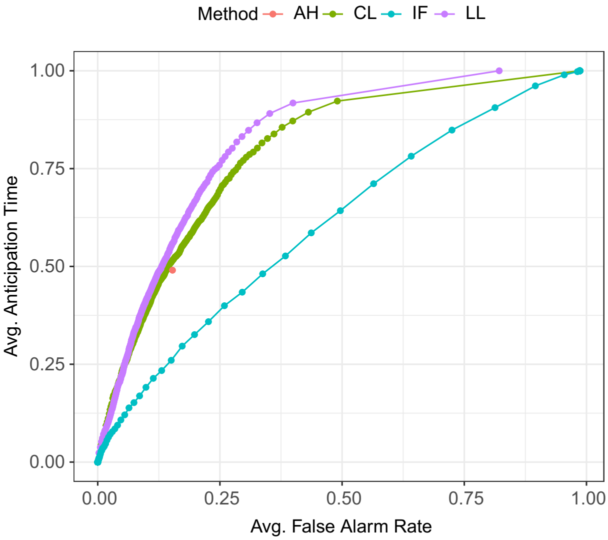


Fig. 6 AMOC curves for the hypertension predictive task

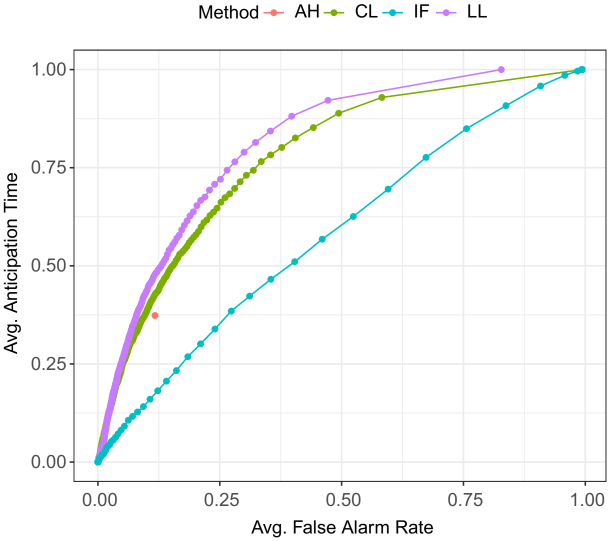


Fig. 7 AMOC curves for the hypotension predictive task

4.5 Results

Figures 4, 5, 6, 7, 8, 9, and 10 show the results of the experiments in the form of AMOC curves, one for each predictive task: bradycardia, bradypena, hypertension, hypotension, hypoxia, tachycardia, and tachypena. Each AMOC curve depicts the false alarm rate (averaged across ICU visits) in the x-axis, and the event anticipation time (also averaged across ICU stays) in the y-axis for the respective approach. The average false alarm rate should be minimized, while the average anticipation time should be maximized. Therefore, methods perform better as their AMOC curve is closer to the top-left corner of the graphic.

Fig. 8 AMOC curves for the hypoxia predictive task

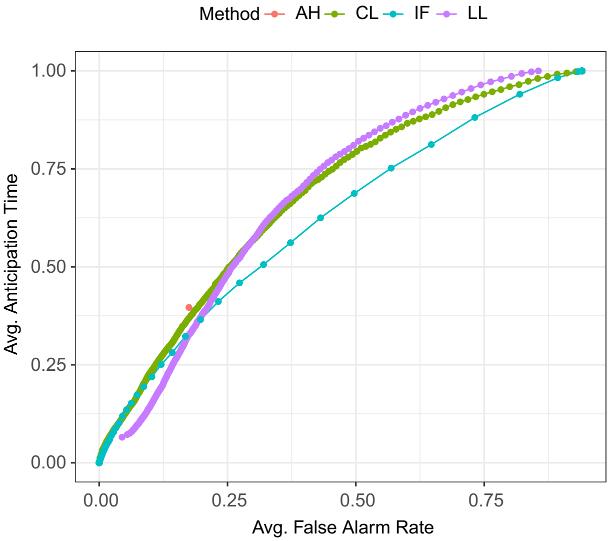
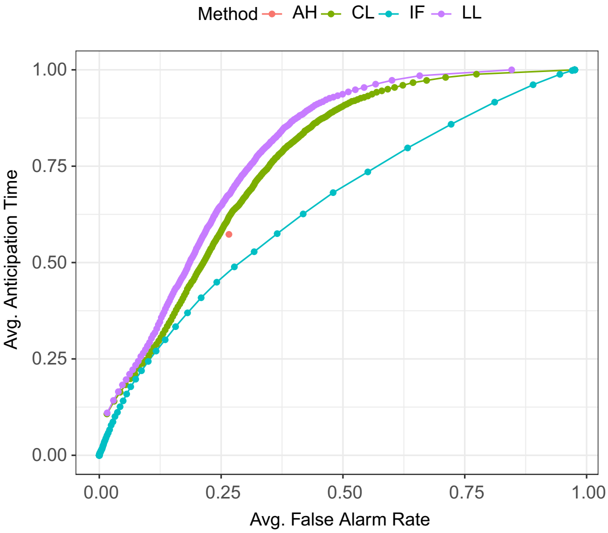


Fig. 9 AMOC curves for the tachycardia predictive task



Inspecting Figs. 4, 5, 6, 7, 8, 9, and 10 from an high-level perspective, the AMOC curve for the proposed method (LL) dominates the curves of the other approaches in most of the problems. In general, both LL and CL show a systematic and consistent advantage relative to IF, except for the tachypena scenario. The AH point is also dominated in most scenarios, except in the bradypena and hypoxia problems. We remark that the AH method presents a single point as it is not a probabilistic approach. Notwithstanding, a single point performance score is less flexible as practitioners are unable to select the level of risk via the decision threshold. Moreover, in the two scenarios where AH is not dominated, its average anticipation score may be too low for detecting CHEs, where the recall scores are

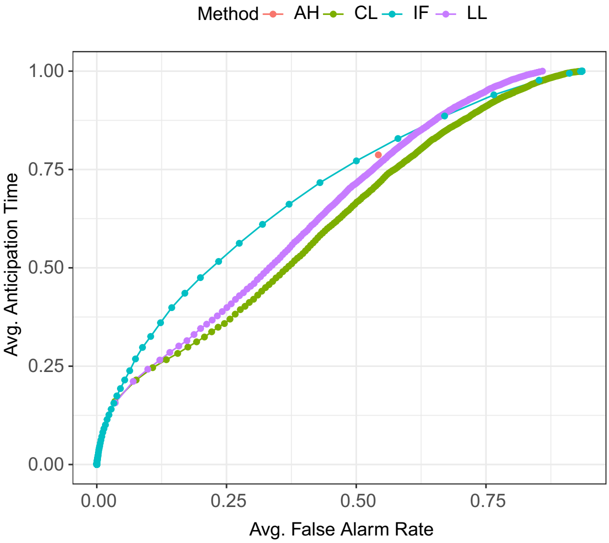


Fig. 10 AMOC curves for the tachypena predictive task

fundamental. It is worth mentioning that `IF` fails to beat the `AH` points in all problems except tachypena, which shows that this method is not generally suitable for these tasks.

`LL` shows a systematic improvement over `CL` in almost all problems. The exceptions are the task hypoxia and bradypena, where the performance of the two methods are comparable across the curves. Interestingly, these are the same two problems where the `AH` point is not dominated by either `LL` or `CL`, tough there may not be any type of causality between the two occurrences.

5 Discussion

5.1 On the experimental results

In the previous section, we provided empirical evidence for the advantages of using a layered learning approach for CHE prediction problems. We briefly discuss the results in this section. We also discuss the main challenges associated with the proposed approach.

The results indicate that `IF`, a state of the art approach for anomaly detection, performs considerably worse relative to discriminating approaches, namely `LL` and `CL`, for six out of the seven problems. While `AH` is a simplistic baseline, its performance was not dominated by the trainable models in two of the tasks. Notwithstanding, this method is also inflexible to different levels of risk. Overall, `LL` shows the most competitive performance relative to state of the art approaches across all seven problems addressed in this work.

As we mentioned before, the reported experiments were carried out using an lightgbm (Ke et al., 2017). This algorithm was used to train both layers of our approach (`LL`), and as a stand-alone classifier without layered learning (`CL`). Using this learning algorithm leads to the best overall results relative to other ones such as random forests, or a deep feedforward neural network. Notwithstanding, deep learning approaches, recurrent architectures in particular, have been increasingly applied in the healthcare domain (e.g., Tamilselvan and Wang (2013)). In future work, we will study these methods further, both as benchmarks and as possible solutions within a layered learning approach.

When applying `CL` and `LL` we resorted to the SMOTE pre-processing algorithm (Chawla et al., 2002) to balance the class distribution. We tested different algorithms, such as random under-sampling, but SMOTE showed the most competitive performance across the seven problems.

In terms of scalability, we did not find significant differences between the trainable methods in terms of inference time. Although `LL` contains two predictive models (one for each layer), the inference time of the lightgbm is negligible.

5.2 Future work

One of the main challenges in the proposed methodology is the manual definition of pre-conditional events. This process is highly domain-dependent. In this sense, it can be regarded as an opportunity for domain experts to embed their expertise in predictive models. Notwithstanding, nowadays there is an increasing interest for end-to-end automated machine learning technologies (Thornton et al., 2013; Feurer et al., 2015), and a manual definition of sub-tasks can be regarded as a bottleneck.

The problem of manually defining sub-tasks is common in other hierarchical approaches. For example, similarly to layered learning, hierarchical reinforcement learning also involves the decomposition of a problem into hierarchical sub-tasks (Dietterich, 2000a). In this topic, one of the most common approaches to this effect is the options framework by Sutton (1998). According to this approach, the definition of the sub-tasks in performed manually by the programmer.

An automatic definition of sub-tasks, which in our case refers to the definition of pre-conditional events, is a difficult problem. In the reinforcement learning literature, there is recent work which try to learn these sub-tasks (Klissarov et al., 2017; Harb et al., 2018; Riemer et al., 2018). In future work, we will explore this research line and try to leverage it to develop a way of automatically defining pre-conditional events. In this work, we settled for simple definition of pre-conditional events by decreasing the target percentage from 90% to 45% for all seven problems. Clearly, this value can be optimized using a validation set, and it is something that an automatic definition of pre-conditional events should include.

Although we focus on CHE prediction problems, our ideas for layered learning can be generally applied to other early anomaly detection or activity monitoring problems, for example, problems with complex targets, which can be decomposed into partial, simpler targets. While the task decomposition is dependent on the domain, we describe some guidelines which can facilitate its implementation.

6 Related research

6.1 Activity monitoring

The problem of timely detection of anomalies is also known in the literature as activity monitoring (Fawcett & Provost, 1999). The goal of this predictive task is to track a given activity over time and launch timely alarms about interesting events that require action.

According to (Fawcett & Provost, 1999), there are two classes of methods for activity monitoring:

- Profiling: In a profiling strategy, a model is constructed using only the normal activity of the data, without reference to abnormal cases. Consequently, an alarm is triggered if the current activity deviates significantly from normal activity. This approach may be useful in complex time-dependent data where anomalies do not have a well-defined concept. For example, fraud attempts often occur in different manners. Effectively, by modelling only normal activity, one is apt to detect different types of anomalies, including the ones unknown hitherto.
- Discriminating: A discriminating method constructs a model about anomalies with respect to the normal activity, handling the problem as a classification one. A system then uses a model to examine the time series and look for anomalies. In this scenario, the recent past dynamics of the data are used as predictor variables. The target variable denotes whether the event of interest occurs.

We focus on the latter strategy, which is the one followed by the proposed layered learning method for activity monitoring. Notwithstanding, we compare our approach to IF, which is a method that follows the profiling strategy.

6.2 CHE prediction

Hypotension prediction has been gaining increasing attention from the scientific community. For example, the 10^{th} annual PhysioNet / Computers in Cardiology Challenge focused on this predictive task (Moody & Lehman, 2009). While the methods used in this particular challenge are not state of the art anymore, the purpose of the reference is to show the relevance of the predictive task.

Regarding the other problems, namely hypertension, tachycardia, bradycardia, tachypena, bradypena, and hypoxia, we follow Forkan et al. (2017) to define these events. However, they adopt a standard classification approach to evaluate the predictive performance of models and check how well each instance is classified into the positive or negative class. In our case, we are interested in anticipating each event rather than each isolated instance. Therefore, we used a more robust approach for defining the critical events. We consider that an event occurs if the respective physiological signal is below/above the threshold in 90% of the observations in a given 30 minute period.

Like other activity monitoring problems or anomaly detection tasks, the typical approach to this problem is to use standard classification methods. This is the case of Lee and Mark, which use a feed-forward neural network as predictive model (Lee & Mark, 2010). Tsur et al. (2018) follow a similar approach and also propose an enhanced feature extraction approach before applying an extreme gradient boosting algorithm.

6.3 Layered learning

Layered learning was proposed by Stone and Veloso (2000), and was specifically designed for scenarios with a complex mapping from inputs to outputs. In particular, they applied this approach to improve several processes in robotic soccer.

Decroos et al. (2017) apply a similar approach for predicting goal events in soccer matches. Instead of directly modelling such events, they first model goal attempts as what we call in this paper as a pre-conditional events sub-task.

Layered learning stems from the more general topic of multi-strategy learning. Layered learning approaches run multiple learning processes to improve the generalisation in a predictive task. This is a similar strategy as ensemble learning methods (Dietterich, 2000b). The main difference is that in layered learning, each layer addresses a different predictive task, while in ensemble learning the predictive task is typically a single one. Another closely related topic to layered learning is hierarchical classification. In hierarchical classification, the output classes have a subsumptive relation. In our case, each pre-conditional event incorporates the occurrence of the respective main event (a CHE). We refer to the works by Silla and Freitas (2011) and Babbar (2014) for a comprehensive read on hierarchical classification approaches. We are unaware of the application of these approaches to time series data, particularly with the proposed idea of pre-conditional events.

6.4 Related early decision systems

The need for early predictions is also important in other predictive tasks which are related to activity monitoring. Time series classification is a well-studied topic, for example, in data stream mining (Bifet & Kirkby, 2009). However, traditional time series classification methods are inflexible for early classification. Typically, a method is trained on the full length of the time series, and the prediction is also made at that time point. Therefore, the main limitation of such methods is that they ignore the sequential nature of data, and the importance of *early* classification (Fawcett & Provost, 1999). The earliness component of classifiers for time series is important so that professionals and decision-makers can take pro-active measures and timely decisions. To overcome this limitation, several models for early classification of time series have been proposed. An example is the work of He et al. (2015), or Xing et al. (2011). Another example of a type of early decision systems is human motion recognition (Kuehne et al., 2011). This task is fundamental for surveillance systems or human-computer interactive systems.

7 Summary

Layered learning approaches are designed to solve predictive tasks in which a direct mapping from inputs to outputs is difficult. In this paper, we developed a layered learning approach for the early detection of anomalies in time series data. The idea is to break the original predictive task into two simpler predictive tasks, which are, in principle, easier to solve. We create an initial model that is designed to distinguish normal activity from a relaxed version of anomalous behaviour (pre-conditional events). A subsequent model is created to distinguish such pre-conditional events from the actual events of interest.

We have focused on predicting critical health conditions in ICUs. Compared to standard classification, which is a common solution to this type of predictive tasks, the proposed model can achieve a better trade-off between average anticipation time and average false alarm rate. The results also suggest that the proposed approach is better than other state of the art methods.

Acknowledgements The authors acknowledge the valuable input of anonymous reviewers.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. et al. (2016) Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283

Babbar, R. (2014) Machine learning strategies for large-scale taxonomies. Ph.D. thesis, Université de Grenoble

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, *77*(2), 81–97.

Bifet, A., & Kirkby, R. (2009) Data stream mining a practical approach. Citeseer

Cao, H., Eshelman, L., Chbat, N., Nielsen, L., Gross, B., & Saeed, M. (2008). Predicting icu hemodynamic instability using continuous multiparameter trends. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (pp. 3803–3806). IEEE

Cerqueira, V., Torgo, L., & Soares, C. (2019) Layered learning for early anomaly detection: Predicting critical health episodes. In *International Conference on Discovery Science*, (pp. 445–459). Springer

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015) xgboost: Extreme gradient boosting, 2017. R package version 0.6-4

Decroos, T., Dzyuba, V., Van Haaren, J., & Davis, J. (2017) Predicting soccer highlights from spatio-temporal match event streams. In *AAAI*, (pp. 1302–1308)

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research, 13*, 227–303.

Dietterich, T. G., et al. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems, 1857*, 1–15.

Fawcett, T., & Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 53–62). ACM

Ferreira, C., Gama, J., Matias, L., Botterud, A., & Wang, J. (2011). A survey on wind power ramp forecasting. Tech. rep., Argonne National Lab.(ANL), Argonne, IL (United States)

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, (pp. 2962–2970)

Forkan, A. R. M., Khalil, I., & Atiquzzaman, M. (2017). Visibid: A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. *Computer Networks, 113*, 244–257.

Ghosh, S., Feng, M., Nguyen, H., & Li, J. (2016). Hypotension risk prediction via sequential contrast patterns of icu blood pressure. *IEEE Journal of Biomedical and Health Informatics, 20*(5), 1416–1426.

Harb, J., Bacon, P.L., Klissarov, M., & Precup, D. (2018). When waiting is not an option: Learning options with a deliberation cost. In *Thirty-Second AAAI Conference on Artificial Intelligence*

He, G., Duan, Y., Peng, R., Jing, X., Qian, T., & Wang, L. (2015). Early classification on multivariate time series. *Neurocomputing, 149*, 777–787.

He, H., & Ma, Y. (2013) Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software, 11*(9), 1–20.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.

Klissarov, M., Bacon, P.L., Harb, J., & Precup, D. (2017). Learnings options end-to-end for continuous action tasks. arXiv preprint arXiv:1712.00004

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011) Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, (pp. 2556–2563). IEEE

Lee, J., & Mark, R. (2010) A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. In *Computing in Cardiology*, 2010, (pp. 81–84). IEEE

Lee, J., & Mark, R. G. (2010). An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *Biomedical Engineering Online, 9*(1), 62.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD), 6*(1), 3.

Moody, G. B., & Lehman, L. W. H. (2009). Predicting acute hypotensive episodes: The 10th annual physionet/computers in cardiology challenge. *Computers in Cardiology, 36*(5445351), 541.

Percival, D.B., & Walden, A.T. (2006) Wavelet methods for time series analysis, (vol. 4). Cambridge university press

Ribeiro, R. P., Pereira, P., & Gama, J. (2016). Sequential anomalies: a study in the railway industry. *Machine Learning, 105*(1), 127–153.

Riemer, M., Liu, M., & Tesauro, G. (2018) Learning abstract options. In *Advances in Neural Information Processing Systems*, (pp. 10,424–10,434)

Saeed, M., Lieu, C., Raber, G., & Mark, R.G. (2002). Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in Cardiology*, 2002, (pp. 641–644). IEEE

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery, 22*(1), 31–72.

Stone, P., Veloso, M. (2000). Layered learning. In *European Conference on Machine Learning*, (pp. 369–381). Springer

Sutton, R.S. (1998). Between mdps and semi-mdps: Learning, planning, and representing knowledge at multiple temporal scales

Tamilselvan, P., & Wang, P. (2013). Failure diagnosis using deep belief learning based health state classification. *Reliability Engineering & System Safety, 115*, 124–135.

Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K. (2013). Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 847–855). ACM

Tsur, E., Last, M., Garcia, V.F., Udassin, R., Klein, M., & Brotfain, E. (2018). Hypotensive episode prediction in icus via observation window splitting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (pp. 472–487). Springer

Weiss, G.M., & Hirsh, H. (1998) Learning to predict rare events in event sequences. In *KDD*, (pp. 359–363) (1998)

Wright, M.N. (2015). ranger: A Fast Implementation of Random Forests. R package

Xing, Z., Pei, J., Yu, P.S., & Wang, K. (2011). Extracting interpretable features for early classification on time series. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, (pp. 247–258). SIAM

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