Layered Learning for Early Anomaly Detection: Predicting Critical Health Episodes

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

[1] University of Porto, Portugal
[2] Dalhousie University, Canada
[3] LIAAD-INESCTEC, Porto, Portugal
vitor.cerqueira@fe.up.pt

Abstract. Critical health events represent a relevant cause of mortality in intensive care units of hospitals, and their timely prediction has been gaining increasing attention. This problem is an instance of the more general predictive task of early anomaly detection in time series data. One of the most common approaches to solve this problem is to use standard classification methods. In this paper we propose a novel method that uses a layered learning architecture to solve early anomaly detection problems. 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 layers, which we hypothesize are easier to solve. Focusing on critical health episodes, the results suggest that the proposed approach is advantageous relative to state of the art approaches for early anomaly detection. Although we focus on a particular case study, the proposed method is generalizable to other domains.

Keywords: time series · early anomaly detection · healthcare · layered learning

1 Introduction

Healthcare is one of the domains which has witnessed a significant growth in the application of machine learning approaches [1]. For instance, intensive care units (ICUs) evolved considerably in recent years due to technological advances such as the widespread adoption of bio-sensors [16]. 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 [8] (AHE) or tachycardia episode [7] (TE) prediction problems. CHEs such as these remain a significant mortality risk factors in ICUs [8], and their timely anticipation is fundamental for improving healthcare.

AHE or TE prediction can be regarded as a particular instance of early anomaly detection in time series data. Fawcett and Provost designated this kind of prediction tasks as *activity monitoring* [5]. Essentially, the goal behind these problems is to issue accurate and timely alarms about interesting future events requiring action.

One of the most common ways to address early anomaly detection problems is to view them as conditional probability estimation problems [5, 19]. Standard supervised learning classification methods can be used for that purpose. The idea is to approximate a function f that maps a set of input observations $X = (x_1, x_2, \ldots, x_n)$ to a binary variable y, which represents whether an anomaly occurs or not. In the case of CHE prediction, the predictor variables (X) summarize 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.

In many domains of application, the anomaly or event of interest is defined according to some rule derived from the data by professionals. For example in healthcare, CHEs are often defined as events where the value of some physiological signal exceeds a pre-defined threshold. Similar approaches for formalizing anomalies can be found in predictive maintenance [14], or wind power prediction [6]. 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 occur simultaneously 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) occurs simultaneously with anomalies (i.e. there can not be an anomaly without a pre-conditional event). This concept is illustrated in the right side of Figure 1 as a Venn diagram for classes.

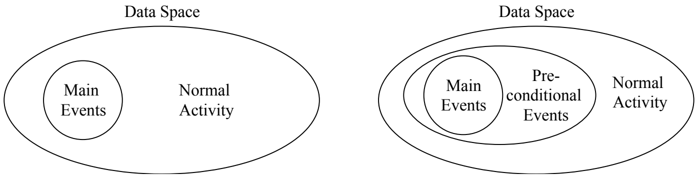

Fig. 1. Venn diagram for the classes in an early anomaly detection problem. The main event represents a small part of the data space; pre-conditional events are more frequent and include the occurrence of the main events.

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 methodology [17]. Layered learning denotes a 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.

We propose a layered learning method to address early anomaly detection problems by splitting the predictive task in two layers (c.f. right side of Figure 1). We first model pre-conditional events relative to normal activity. A subsequent model is applied to distinguish pre-conditional events from the actual anomalies.

Effectively, the first layer affects the learning process of the second layer by decreasing the scope of its data space. Since the model in the second layer is created to distinguish the events of interest from pre-conditional events, it does not train on observations of what is designated as normal activity.

We apply the proposed approach to tackle the problem of CHE prediction. In the experiments, the layered learning model shows a better predictive performance relative to state of the art approaches, including a direct classification approach (without layered learning, see the left side of Figure 1). In short, the contributions of this paper are:

- the idea of pre-conditional events in time series;
- a general layered learning approach to the early detection of events in time series data;
- the application of the proposed approach to AHE and TE prediction.

All work and results presented in the paper are reproducible. The data is publicly available [16], and the code for the methods can be found at `https://github.com/vcerqueira/layered_learning_time_series`.

2 Early Anomaly Detection in Time Series

Let $\mathcal{E} = \{E_1, \ldots, E_{|\mathcal{E}|}\}$ denote a set of time series. For example, $\mathcal{E}$ may represent a set of patients being monitored at the ICU of an hospital. Each time series $E \in \mathcal{E}$ can be represented as a set of subsequences $E = \{e_1, e_2, \ldots, e_i, \ldots, e_{t-1}, e_t\}$, where e_i represents the *i-th* subsequence. A subsequence denotes a tuple $e_i = (t_i, X_i, y_i)$, where t_i denotes the time stamp that marks the beginning of the subsequence, $X_i \in \mathbb{X}$ represents the input (predictor) variables, which summarize the recent past dynamics of the time series; and $y_i \in \mathbb{Y}$ denotes the target variable, which is a binary value ($y_i \in \{0, 1\}, \forall\ i \in \{1, \ldots, t\}$) that represents whether or not there is an impending anomaly or event of interest in the near future of the respective time series. How near in the future is typically a domain-dependent parameter.

For each subsequence e_i we construct the feature-target pair (X_i,y_i) as follows. As illustrated in Figure 2, each subsequence 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

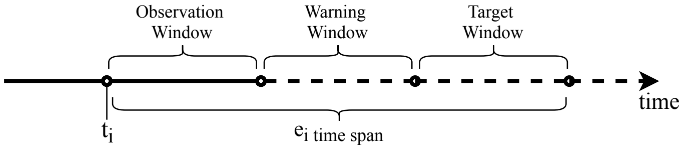

Fig. 2. Splitting a subsequence e_i into observation window, warning window, and target window. The features X_i are computed during the observation window, while the outcome y_i is determined in the target window.

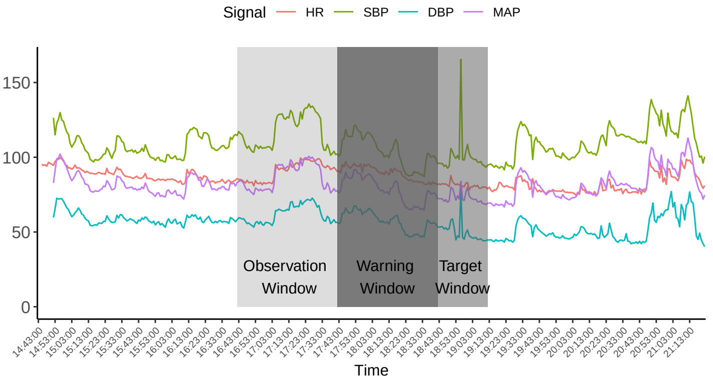

Fig. 3. The physiological signal of patients are monitored over time. Each subsequence, denoted by the shaded areas, is split in an OW, a WW, and a TW.

values of X_i; and (iii) a warning window (WW), which is the lead time necessary for a prediction to be useful. For instance, in clinical medicine physicians need some time after an alarm is launched, for example to decide the most appropriate treatment. The sizes of these windows are domain-dependent. In principle, the problem will be easier as the OW is closer to the TW, that is, a smaller WW is required [12, 20].

2.1 Event Prediction in ICUs

In this paper we focus on a particular instance of early anomaly detection problems: CHE prediction in ICUs, namely AHE and TE. Ghosh et al. [8] state that prolonged hypotension leads to a critical health damage, from cellular dysfunction to severe injuries in multiple organs. In turn, sustained tachycardia significantly increases the risk of stroke or cardiac arrest.

Patients assigned to the ICU are typically monitored constantly, with biosensors capturing several physiological signals, such as heart rate, or mean arterial blood pressure. This is illustrated in Figure 3, where the data of a patient is depicted. A subsequence for CHE prediction is given as example in the shaded area of the graphic.

Acute Hypotensive Episodes. Hypotension episodes denote a prolonged drop in the blood pressure. More formally, AHE is an event defined as "a 30-minute window having at least 90% of its mean arterial blood pressure (MAP) values below 60 mmHg [millimeters of mercury]" [12, 19]. In this context, the target variable value is computed as follows:

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

In other words, we consider that the i-th subsequence represents an anomaly if its TW represents an AHE (c.f. Figure 3). Since AHEs are rare, the target vector y is dominated by the negative class (i.e. $y = 0$), where a patient shows a normotensive status. For the target window of 30 minutes, we consider an OW and a WW of 60 minutes each. These values are typically used in the literature of AHE prediction models [8].

Tachycardia Episodes. Tachycardia denotes a high heart rate (HR). Generally, an HR over 100 beats per minute (bpm) under a resting state is considered as tachycardia. In order to consider a more robust definition for the purpose of discovering tachycardia episodes we follow a similar intuition to AHEs. We define TE as "a 30-minute window having at least 90% of its HR values above 100 bpm". The respective target variable is computed as follows:

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

TEs are defined similarly to AHEs. Moreover, TEs also denote rare events since ICU patients usually show a HR below 100 bpm. We consider identical window sizes (OW, WW, TW) for both problems.

2.2 Discriminating approaches to early anomaly detection

Naturally, one of the most common approaches to solve the problem defined previously is to view it as a conditional probability estimation problem and use standard supervised learning classification methods [5, 19]. The idea is to build a model $f : \mathbb{X} \rightarrow \mathbb{Y}$, which 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 (c.f. left side Figure 1).

Notwithstanding the widespread 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. In essence, layered learning consists in breaking a predictive task into several layers. The approach assumes that the problem addressed in each layer is simpler than the original one. As Stone and Veloso explain, “the key defining characteristic of layered learning is that each layer directly affects the learning of the next” [17]. 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.

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, an AHE is defined as a percentage of numeric values within a time interval which are below some threshold (c.f. Section 2.1). TEs are defined in a similar manner. This type of approach for defining anomalous events is also common in other domains. For example in predictive maintenance [14], where numerical information from sensor readings is transformed into a class label which denotes whether or not an observation is anomalous. Or wind ramp detection, where a ramp event is a rare occurrence that denotes a large percentual change in wind power in a short time interval [6].

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 simultaneously occur with the main event taking place. 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 AHE prediction. In Section 2.1, we defined the main event (AHE) 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”. In summary, pre-conditional events should have the following two characteristics: (i) pre-conditional events should have a higher relative frequency than the main events; and (ii) pre-conditional events always happen when the main events happen. The inverse is not a necessary condition.

3.3 Our Approach

We can leverage the idea of pre-conditional events and use a layered learning strategy to tackle early anomaly detection problems in time series data. Our idea is to decompose the main predictive task into two layers, each denoting a predictive subtask. 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 Figure 1. The figure presents two Venn diagrams for classes. Focusing on the left-hand side, the anomalies or main events (e.g. AHE) 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 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 early anomaly detection tasks. This phenomenon can compromise the performance of learning algorithms [9]. In effect, we first model the pre-conditional events with respect to the normal activity. These pre-conditional events are, in principle, easier to learn relative to the main concept as 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.

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

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

For this task a subsequence $e_i^{\mathcal{S}}$ is a tuple $e_i^{\mathcal{S}} = (t_i, X_i, y_i^{\mathcal{S}})$. The difference to the original set of subsequences E 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 variables X to the output $y^{\mathcal{S}}$.

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 subsequence". Effectively, the target variable for this task is defined as follows:

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

The target variable for this subtask ($y^{\mathcal{F}}$) is formalized 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 subsequence $e_i^{\mathcal{F}}$ is represented by $e_i^{\mathcal{F}} = (t_i, X_i, y_i^{\mathcal{F}})$. In this scenario however, the set of available subsequences E is considerably less than in the pre-conditional sub-task because only the subsequences for which $y^{\mathcal{S}}$ equals 1 are accounted for. 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.

Forecasting 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 subsequence e_i according to the multiplication of the outcome predicted by both $f^{\mathcal{S}}$ and $f^{\mathcal{F}}$.

3.4 Application of Layered Learning to CHE Prediction

As mentioned before (c.f. Section 2.1) an AHE is defined as a 30-min time period where 90% of the blood pressure values are below 60 mmHg. We propose to relax this threshold and define the pre-conditional event event $\mathcal{S}$ as follows. We define $\mathcal{S}^{\text{AHE}}$ to represent "a 30-minute window having at least **45**% of its mean arterial blood pressure values below 60 mmHg".

The event $\mathcal{S}$ is consistent with the two above-mentioned characteristics: the frequency of $\mathcal{S}$ across the database is considerably higher than an AHE – note that the blood pressure level can drop below 60 mmHg for some time period without being considered as an hypotensive episode. Consequently, the occurrence $\mathcal{S}$ is simultaneous to the occurrence of an AHE (if 90% of the values are below 60 mmHg, so are 45%).

We apply the same reasoning to the TE prediction task. In Section 2.1, we defined a TE as "a 30-minute window having at least 90% of its HR values above 100 bpm". In order to define $\mathcal{S}^{\text{TE}}$ we again relaxate the percentage threshold as follows. $\mathcal{S}^{\text{TE}}$ is defined as "a 30-minute window having at least 45% of its HR values above 100 bpm". In both situations, the value of 45% was chosen arbitrarily. We attempted to make the pre-conditional events much more frequent relative to the main events. Nevertheless, this parameter can be optimized.

4 Empirical Evaluation

4.1 Case Study: MIMIC II

In the experiments we used the database Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC) II [16], which is a benchmark for a number of predictive tasks in healthcare, including CHE prediction.

As inclusion criteria of patients and general database pre-processing steps, we follow Lee and Mark closely [12]. For example, the sampling frequency of the physiological data of each patient in the database is one minute. Moreover, the following physiological signals are collected: heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial blood pressure (MAP). As described in Section 2.1, the TW size is 30 minutes. For each TW, there is a 60-minute OW and a 60-minute WW. For a comprehensive read regarding the data compilation we refer to the work by Lee and Mark [12]. Considering this setup, the number of patients is 1,072, leading to a data size of 1,975,936 subsequences. 71,035 of those events represent an AHE (about 3.5%). In turn, 13.6% of the subsequences represent a TE.

Regarding feature engineering, we follow previous work in the literature [11,19]. Using the observation window of each subsequence 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 carry a wavelet transformation to capture the relative energies in different spectral bands.

4.2 Experimental Design

The experiments were designed to compare the proposed layered learning approach to state of the art methods for early anomaly detection. To estimate the predictive performance of each method we used a 5×10-fold cross-validation, in which folds are split by patient. To be more precise, in each iteration of the cross-validation procedure, one fold of the set of patients $\mathcal{E}$ is used for validation, another fold of different patients is used for testing, and the remaining patients are used for training the predictive model. The set of time series $\mathcal{E}$ only comprises a temporal dependency within each patient, and we assume the data across patients to be independent. In this context, the application of cross-validation in this setting is valid. Finally, the subsequences of the patients chosen for training are concatenated together to fit the predictive model.

The goal behind early anomaly detection problems is not to classify each subsequence as positive or negative [5]. Instead, the main goal is to detect, in a timely manner, when there is an impending anomalous event. In this context, we follow Weiss and Hirsh [20] regarding the evaluation metrics. Specifically, two measures are computed: Event Recall (ER), and Reduced Precision (RP). These two metrics follow the same intuition of the widely used Recall and Precision metrics, but are tailored for time-dependent data.

Let T denote the total number of events of interest in a test data set, and let $\hat{T}_m$ represent the total number of those events correctly predicted by a model m. The ER for model m is given by the following equation:

$$\mathrm{ER}_m = \frac{\hat{T}_m}{T} \tag{4}$$

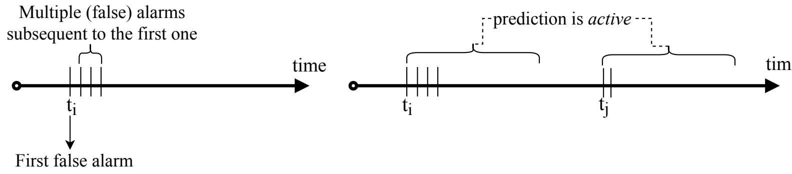

Fig. 4. Left: A sequence of consecutive false alarms – the first alarm is useful, but the subsequent ones may add no information; right: false alarms (denoted as vertical bars) over a time interval – there are 6 false positives, but only two discounted false positives.

ER differs from the classical recall metric because a single correct prediction within an observation window leading to an event is enough to consider that event correctly anticipated. As Fawcett and Provost put it, "alarming earlier may be more beneficial, but after the first alarm, a second alarm on the same event may add no value" [5].

The classical precision metric measures the percentage of positive predictions that are correct. Similarly to recall, in a time-dependent domain the classical precision may be misleading because multiple predictions on the same event are counted multiple times. This idea is intuited in Figure 4 (left). This graphic shows a subsequence in which predictions are being produced over time. Starting from time t_i, four false alarms are triggered. Performance evaluation should take the first wrong prediction into account as a false positive. However, the subsequent false alarms (as shown in the left side of Figure 4) are not meaningful since they add no information – assuming some action is taken after the first alarm.

RP overcomes this problem by considering a prediction to be *active* for some time period. Specifically, in this work we consider a time interval with the same size as the observation window. Notwithstanding, this is usually a domain dependent parameter. Effectively, the RP metric replaces the number of false positives with the number of discounted false positives – the number of non-overlapping observation periods associated with a false prediction. This idea is illustrated in Figure 4 (right), where each vertical bar in the time line denotes an issued false alarm. There are a total of 6 false positives, but, if taking into account the time interval a prediction is active, there are only two discounted false positives (DFP). Finally, RP also considers the number of target events correctly identified ($\hat{T}_m$), instead of the number of correct predictions (true positives). In effect, RP for model m is given by the following equation:

$$\mathrm{RP}_m = \frac{\hat{T}_m}{\hat{T}_m + \mathrm{DFP}_m} \tag{5}$$

4.3 Learning Algorithm and State of the Art Methods

In the experiments we tested different predictive models, namely a random forest, a support vector machine, a deep feed-forward neural network, and an extreme gradient boosting model [3]. We only show the results of the latter in these experiments, since it provides a better performance than the remaining methods for both AHE prediction and TE prediction.

The classifiers used in the experiments output a probability. The decision threshold is optimized following previous work in the literature of AHE prediction [12, 19], which recommends selecting the threshold that maximizes the average of *classical* recall and specificity (true negative rate).

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

- `CL` a standard classification method that does not apply a layered learning approach and directly models the events of interest with respect to normal activity (c.f. Figure 1) – in order to cope with the class imbalance problem this approach includes a random under-sampling of the majority class;
- `IF` the Isolation Forest [13] method, which is a state of the art model-based approach to anomaly detection;
- `RG` We include a regression-based alternative both for AHE prediction and TE prediction [12, 15]. We apply a multi-step forecasting model to predict the future values of MAP (for AHEs) and HR (for TEs). Regarding the former, and following up on the definition of an AHE (Section 2.1), an alarm for an AHE is triggered if 90% of the forecasted values for the MAP variable are below 60 mmHg [15]. Likewise, an alarm for an TE is triggered if 90% of the forecasted values for the HR variable are above 100 bpm. The multi-step forecasting model follows a *direct* approach [18];
- `AH` 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 AHE 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 TE prediction, where an alarm is launched if the HR variable exceeds 100 bpm.

4.4 Results

Table 1 presents the average results, and respective standard deviation, for each method across the 50 folds ($5 \times 10-$fold cross-validation). We analyse the significance of the results according to the Bayesian correlated t-test [2] (Figures 5 and 6). In the Bayesian correlated t-test we consider the region of practical equivalence to be the interval $[-0.01, 0.01]$. In other words, two methods are practically equivalent if their difference in performance is below 0.01.

In terms of ER, on average the proposed method `LL` captures 83% of AHEs and 92.5% of TEs. These values are significantly better relative to the remaining methods, including `CL` which is the typical approach to solve these predictive tasks. Maximizing ER in this particular domain of application is important, because the events of interest are disruptive. Regarding RP, overall the value of this metric is generally low for all methods, which suggests an high number

Table 1. Average of results for the CHE prediction problem across the 50 folds

	AHE		TE	
Method	**ER**	**RP**	**ER**	**RP**
AH	0.625±0.05	0.129±0.02	0.749±0.05	**0.204**±0.02
CL	0.794±0.05	0.095±0.01	0.909±0.03	0.135±0.02
IF	0.700±0.18	0.035±0.01	0.756±0.31	0.051±0.01
LL	**0.830**±0.05	0.090±0.01	**0.925**±0.02	0.140±0.01
RG	0.250±0.06	**0.205**±0.04	0.646±0.04	0.195±0.02

of DFP. In relative terms, AH and RG show the best results on this metric. The proposed approach (LL) shows a comparable RP with CL. Expectedly, there is a trade-off between ER and RP: greater ER leads to lower RP, and vice-versa. Notwithstanding, relative to CL, LL is able to significantly improve ER while keeping a comparable (i.e., within the region of practical equivalence) RP. While LL shows a significantly worse RP relative to AH and RG, it compensates with a considerably better ER. In comparison with IF, LL is significantly better in both metrics.

4.5 Discussion

In the experiments above we showed the competitiveness of the proposed method for early anomaly detection in a case study from the healthcare domain. The main challenge behind layered learning is the assumption that the task decomposition is a domain-dependent function. This can be regarded as an opportunity for domain experts to embed their domain expertise in predictive models. Notwithstanding, nowadays there is an increasing interest for end-to-end automated machine learning technologies, and a manual decomposition can be

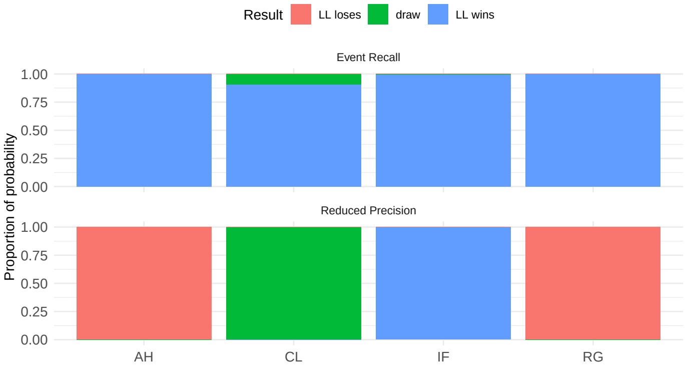

Fig. 5. Comparing CL with LL with a Bayesian correlated t-test for ER and RP metrics (AHE prediction)

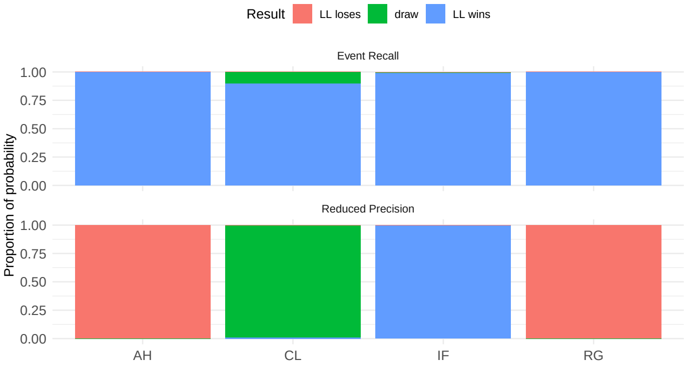

Fig. 6. Comparing CL with LL with a Bayesian correlated t-test for ER and RP metrics (TE prediction)

regarded as a bottleneck. In this context, future work includes the study of an automated methodology for identifying or learning the pre-conditional events from the data.

Although we focus on CHE prediction problems, our ideas for layered learning can be generally applied to other early anomaly detection 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.

5 Related Work

5.1 Early Anomaly Detection and CHE Prediction

According to Fawcett and Provost [5], there are two classes of methods for activity monitoring: profiling methods, and discriminating methods. 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 the normal activity. On the other hand, 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. We focus on the latter strategy, which is the one followed by the proposed layered learning method for early anomaly detection. Notwithstanding, we compare our approach to `IF`, which is a method that follows the profiling strategy.

Like other early anomaly detection problems, the typical approach to tackle CHE prediction problems is to use standard classification methods. This is the case of Lee and Mark, which use a feed-forward neural network as predictive algorithm [12]. Tsur et al. follow a similar approach, and also propose an enhanced feature extraction approach before applying an extreme gradient boosting algorithm [19]. In turn, Rocha et al. propose a regression approach (`RG`) by forecasting future values of blood pressure [15]. In their approach, alarms for impending AHE are launched according to a deterministic function which receives as input the numeric predictions. TE prediction also is a relevant task. For example, Forkan et al. [7] propose a predictive model for detecting several health conditions, including tachycardia and hypotension.

5.2 Layered Learning

Layered learning was proposed by Stone and Veloso, and was specifically designed for scenarios with a complex mapping from inputs to outputs [17]. In particular, they applied this approach to improve several processes in robotic soccer. Decroos et al. [4] 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 surrogate task. Layered learning is part of the family of hierarchical models. To our knowledge, this is the first time such an approach is applied to early anomaly detection using the idea of pre-conditional events.

6 Final Remarks

In this paper we developed a layered learning approach for the early detection of anomalies in time series data. We create an initial model that is designed to distinguish normal activity from a relaxed version of anomalous behavior (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 is able to capture significantly more anomalous events with a comparable number of false alarms.

Future work includes: (i) a better understanding of how layered learning works, how to tune its parameters; (ii) its application to tackle other early anomaly detection problems; (iii) automatic identification of pre-conditional events– we are studying the usage of subgroup discovery [10] to this effect.

Acknowledgements

Vitor Cerqueira is supported by a FCT PhD research grant (SFRH/BD/135705/2018).

References

1. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics **77**(2), 81–97 (2008)

2. Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. The Journal of Machine Learning Research **18**(1), 2653–2688 (2017)
3. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: xgboost: Extreme gradient boosting, 2017. R package version 0.6-4 (2015)
4. Decroos, T., Dzyuba, V., Van Haaren, J., Davis, J.: Predicting soccer highlights from spatio-temporal match event streams. In: AAAI. pp. 1302–1308 (2017)
5. Fawcett, T., Provost, F.: 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 (1999)
6. Ferreira, C., Gama, J., Matias, L., Botterud, A., Wang, J.: A survey on wind power ramp forecasting. Tech. rep., Argonne National Lab.(ANL), Argonne, IL (United States) (2011)
7. Forkan, A.R.M., Khalil, I., Atiquzzaman, M.: 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 (2017)
8. Ghosh, S., Feng, M., Nguyen, H., Li, J.: Hypotension risk prediction via sequential contrast patterns of icu blood pressure. IEEE journal of biomedical and health informatics **20**(5), 1416–1426 (2016)
9. He, H., Ma, Y.: Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons (2013)
10. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with cn2-sd. Journal of Machine Learning Research **5**(Feb), 153–188 (2004)
11. Lee, J., Mark, R.: 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 (2010)
12. Lee, J., Mark, R.G.: An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. Biomedical engineering online **9**(1), 62 (2010)
13. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(1), 3 (2012)
14. Ribeiro, R.P., Pereira, P., Gama, J.: Sequential anomalies: a study in the railway industry. Machine Learning **105**(1), 127–153 (2016)
15. Rocha, T., Paredes, S., De Carvalho, P., Henriques, J.: Prediction of acute hypotensive episodes by means of neural network multi-models. Computers in biology and medicine **41**(10), 881–890 (2011)
16. Saeed, M., Lieu, C., Raber, G., Mark, R.G.: Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In: Computers in Cardiology, 2002. pp. 641–644. IEEE (2002)
17. Stone, P., Veloso, M.: Layered learning. In: European Conference on Machine Learning. pp. 369–381. Springer (2000)
18. Taieb, S.B., Bontempi, G., Atiya, A.F., Sorjamaa, A.: A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. Expert systems with applications **39**(8), 7067–7083 (2012)
19. Tsur, E., Last, M., Garcia, V.F., Udassin, R., Klein, M., Brotfain, E.: 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 (2018)
20. Weiss, G.M., Hirsh, H.: Learning to predict rare events in event sequences. In: KDD. pp. 359–363 (1998)