Hypotensive Episode Prediction in ICUs via Observation Window Splitting*

Elad Tsur[1], Mark Last[1], Victor F Garcia[2], and Raphael Udassin[3]

[1] Dept. of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel. eladtsur@gmail.com, mlast@bgu.ac.il
[2] Division of Pediatric Surgery, MLC 2023, Children's Hospital Medical Center, 3333 Burnet Ave., Cincinnati, Ohio 45229, USA. victor.garcia@cchmc.org
[3] Pediatric Surgery Department, Hadassah University Hospital, Ein-Karem, Jerusalem 9112001, Israel. raphaelu@ekmd.huji.ac.il

Abstract. Hypotension, defined as dangerously low blood pressure, is a significant risk factor in intensive care units (ICUs), which requires prompt therapeutic intervention. The goal of our research is to predict a Hypotensive Episode (HE) by time series analysis of continuously monitored physiological vital signs. Particularly, we aim to give the physicians a one-hour warning using a prognostic model based on the last Observation Window (OW) at the prediction time. Existing clinical episode prediction studies used a single OW of 5–120 minutes to extract predictive features, where no significant improvement was reported when longer OWs were used. In this work we have developed the *In-Window Segmentation* (InWiSe) method for time series prediction, which splits a single OW into several sub-windows of equal size. The resulting feature set combines the features extracted from each observation sub-window, and we then used the Extreme Gradient Boosting (XGBoost) binary classifier to produce an impending episode prediction model from the combined feature set. We evaluate the proposed approach on three retrospective ICU datasets (extracted from MIMIC II, Soroka and Hadassah databases) using cross-validation on each dataset separately, as well as by cross-dataset validation. The results show that InWiSe is superior to existing methods in terms of the area under the ROC curve (AUC).

Keywords: Time series analysis · Clinical episode prediction · Feature extraction · Intensive care · Patient monitoring

1 Introduction

Hypotension is defined as dangerously low blood pressure. It is a major hemodynamic instability symptom, as well as a significant risk factor in hospital mortality at intensive care units (ICUs) [1]. As a critical condition which may result in

* Partially supported by the Cincinnati Children's Hospital Medical Center; In collaboration with Soroka Medical Center in Beer-Sheva and Hadassah University Hospital, Ein Karem, Jerusalem

a fatal deterioration, an impending Hypotensive Episode (HE) requires prompt therapeutic intervention [2] by ICU clinicians. However, HE prediction is a challenging task for several reasons [3]. First, the amount of time which can be allocated by the clinical staff per patient is limited. Second, the amount of accumulated physiologic data per patient is massive in terms of both data variety (multi-channel waveforms, laboratory results, medication records, nursing notes, etc.) and data volume (length of waveform time series). Last, even with sufficient time, resources, and data, it is very hard to induce an accurate estimation of the likelihood of clinical deterioration with bare-eye analysis alone.

HE may be detectable in advance by analysis of continuously monitored physiologic data; more specifically, the analysis of vital signs (multi-parameter temporal vital data), may inform on the underlying dynamics of organs and cardiovascular system functioning. Particularly, vital signs may contain subtle patterns which point to an impending instability [4]. Such pattern identification is a suitable task for machine learning algorithms. Smart patient monitoring software that could predict the clinical deterioration of high risk patients well before there are changes in the parameters displayed by the current patient monitors would save lives, reduce hospitalization costs, and contribute to better patient outcomes [5].

Our research goal is to give the physicians a one hour warning of an impending HE by building a prediction model, which utilizes the maximal amount of information from the currently available patient monitoring data and outperforms current state-of-the-art HE prediction systems.

This paper is organized as follows. Section 2 surveys the previous works in several related areas, elaborates on the limitations of these works and introduces the contributions of our method. Section 3 describes the studied problem and proposed methods in detail and Section 4 covers the results of the empirical evaluation. Finally, Section 5 presents the conclusions along with possible directions for future research.

2 Related Work and Original Contributions

Several works studied the problem of clinical deterioration prediction in ICUs. This section reviews their problem definitions, feature extraction methods, sliding window constellations, and prediction algorithms. Finally, a discussion of the limitations of existing methods is followed by a presentation of the contributions of this study.

2.1 Clinical Episode Definitions

Previous works vary mainly in two aspects of their clinical deterioration prediction problem definition [6]. The first is how to define an episode. For example, the definition could be based on the recorded clinical treatment or on the behavior of vital signs within a specific duration of time. The second is how long before an episode one aims to predict it; this also called the *Gap Window* and will be called the *gap* in this study.

The objective in [3] was to predict the hemodynamic instability start time with a 2-hours gap. The episode start time was defined by a clinical intervention recorded in the ICU clinical record of a patient. In [7], instability was also defined by some given medications, and gaps of 15 minutes to 12 hours were explored.

The 10^{th} annual PhysioNet / Computers in Cardiology Challenge [4] conducted a competition to study an Acute Hypotensive Episode (AHE). They defined AHE as an interval in which at least 90% of the time the Mean Arterial blood Pressure (MAP) is under 60 mmHg during any 30-minute window within the interval. Their goal was to predict whether an AHE will start in the next 60 min. In [1, 8], the HE and AHE definitions were identical to in [4], but a lower MAP bound of 10 mmHg was added to prevent noise effects from outliers. Their goal was to predict the patient condition in a *Target Window* (called herein *target*) of 30 minutes, which occurs within a gap of 1-2 hours (See Figure 1a), and label it as hypotensive or normotensive (normal blood pressure). As expected, and as concluded in [8], the problem is more challenging when predicting further into the future, thus resulting in poorer performance. Note that, as indicated in [8], the accepted HE definitions for adults vary in the range of 60-80 mmHg MAP for 30+ min, where the lowest case of 60 mmHg is sometimes excluded under the definition of AHE [1, 4].

A generalization of clinical deterioration prediction was done in [5], where episode was defined as one of seven critical patient conditions (Tachycardia, Hypertension, etc.) which are all defined by some of four vital signs. The critical condition episode definition was at least 30 minutes in which all four vital signs deviate from their normal range. The tested gaps were 60, 90, and 120 minutes.

In [9], a different approach that does not formally define an episode was used by producing an anomaly-based early warning, which alerts according to parameter fluctuations by some amount of standard deviations.

2.2 Feature Types

In most works, the future episode predictive features are usually extracted from a sliding *Observation Window* (OW) over a record, which is a collection of vital sign time series of one patient in a single ICU admission. Furthermore, a minute-by-minute vital signs time series, like blood pressure and Heart Rate (HR) are usually used to extract features, while a few works used the clinical information (age and temporal medications data) as well. Moreover, the difference between works lies in the type of features extracted from the vital signs, also called parameters; we note that the typically used benchmark database is MIMIC II [10], a multi-parameter ICU waveforms and clinical database.

Some studies start by adding new *knowledge-based* parameters calculated based on the biological and clinical nature of vital signs, as well as their inter-relations. In [3], starting with HR, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP) and MAP, four new parameters are derived (i.e., HR to Blood Pressure Ratio), while in [8] the pulse pressure and relative cardiac output are derived. In an alternative knowledge-based approach [7] fifteen decision rules on medical events and vital signs were selected as predictive indices.

Statistical features are the most obvious source for the extraction of predictive features, also called patterns, from intervals like OWs. In [5], the authors calculate extremes, moments, percentiles and inter-percentile ranges for every vital sign, whereas in [8] interquartile ranges and slope are added. In a more pragmatic statistical approach [11], several episode predictive indices were used, derived from the blood pressure signals only. These indices were six types of averages from SBP, DBP and MAP, each taken as a single feature.

Another statistical approach derives *cross-correlation* features which capture the coupling between two time series by computing the sum of products of their values [8], or by estimating their variance and covariance [5].

A more recent and widely accepted feature extraction approach is the use of *wavelets*, which captures the relative energies in different spectral bands that are localized in both time and frequency. Wavelets were proven to be predictive patterns for time series and perform well as episode predictors [12] or as vital sign similarity detectors [13]. In [5] and [8] Daubechies (DB) and Meyer wavelet types were used, respectively, noting that the DB type dominates the basic Haar type wavelets [14] in terms of vital sign time series, which are non-stationary [5].

In *sequential patterns* mining, time series are searched for templates that repeat with some support over many instances (i.e., OWs). Ghsoh et al. [1, 15] mined blood pressure signals by finding sequential contrast patterns having the most discriminative capabilities in order to predict future HEs and septic shocks.

Apart from vital signs, the age and vasopressors (blood pressure medications) given during OWs are added as features by Lee and Mark [8]. On finding that medication events have low correlation with the target, in their further work [16] they achieve similar results without those features. Moreover, in [12], the unreliability of vasopressors medication recordings is mentioned in terms of timing accuracy, which is very important for this study task.

2.3 Observation Window Constellations

The sliding OW plays an important role in the episode prediction task. In this section, we survey the different approaches of constructing and collecting OWs.

The first important attribute of an OW is its duration. In [1, 3, 5, 8, 11, 15], various OW sizes were applied (5, 10, 30, 60, 90, and 120 minutes). Having implemented a 60-min OW, it is claimed in [1, 8] that extracting features from a longer window does not result in improvement of prediction performance.

In [16], Lee and Mark extended their previous work by implementing a weighted decision classifier that consists of four base classifiers, each predicting an HE in the same target but within different gaps (1, 2, 3 and 4 hours) using a different corresponding 30-min OW. The final decision is made by weighting the four posterior probabilities from each classifier. They report insignificant improvement in prediction performance as well as independency of predictions from the 3^{rd} and the 4^{th} past hours.

A second matter is how to collect OWs for training the prediction algorithm. One simple approach, applied by Cao et al. [3], is for every unstable record (having one or more HEs), to compile an OW ending gap-minutes before the first

episode start time. For each stable record, one or more OWs are then sampled randomly. Following that in [8], collecting multiple OWs from either stable or unstable records (regardless of the episode locations) and in a random fashion which does not collect windows exactly gap-minutes before an episode start is proved to outperform the first method. A sliding target window (with no overlap) traversed each record (Fig 1a), and as many OWs as possible were compiled. However, they note that collecting OWs all the time, even when no HE is impending, and doing it from both stable and unstable patients will result in an extremely imbalanced dataset. Having two OW classes (hypotensive or normotensive), one way to solve this issue is by undersampling the majority class (normotensive) [5, 8].

Previous studies included or excluded OW instances from their dataset according to the OW validity. In [1, 5, 8, 16], a valid OW is defined by having less than 5% outliers or missing values, where an outlier is a value beyond the $10-200$ range for any vital sign.

2.4 Prediction Algorithm

The problem of episode prediction is most logically approached by calculating the probability of a future episode at a certain point, using the current OW, and then classifying the target, which starts within some gap-minutes, as hypotensive or not, based on a pre-defined probability threshold. Multiple works tackled this problem by using numerous supervised machine learning algorithms, particularly binary classifiers, with some exceptions such as in [17], where a Recurrent Neural Networks approach is used to forecast the target window MAP values, followed by a straightforward binary decision based on the episode definition.

The classifiers chosen by some other papers are Logistic Regression [3], Artificial Neural Network [8, 16], Majority Vote [1] and Random Forest [5]. To the best of our knowledge, the most accurate HE prediction so far is that reported in [8, 16], which we reproduce and use for comparison in Section 4.

2.5 Limitations of Published Methods

Advanced methods and evaluation schemes such as in [1, 5, 8, 15, 16], solved some of the problems found in the early works [3, 11], yet left some open issues, including low precision (14% in [8]) and a strict episode definition that is still far from the practical definitions used in ICUs. Moreover, a machine learning solution to a high precision HE prediction will probably need much more training data, while the current MIMIC II [10] contains only several thousands[4] of vital sign records that are matched with the clinical data (necessary for differentiating between children and adult patients who have different episode definitions).

However, there are three other important limitations in the area of this study. First, there is a lack of more comprehensive public sources of ICU monitored vital signs in addition to the existing MIMIC database. The current episode prediction

[4] The MIMIC III waveform database *Matched Subset*, four times larger than MIMIC II matched subset, was published very recently and should be included in future work.

works miss the crucial cross-dataset validation. This should be in the interest of anyone aiming to find a generic model to work for any ICU, which may or may not have retrospective patient records for training.

Second, recent papers [1, 5, 8] include predictions of future episodes even if the patient is during an ongoing episode. This information may be nonessential to physicians on one hand, and the corresponding metrics of such an experiment will not necessarily reflect the performance of an applied system on the other.

Finally, works conducted over the last decade show no improvement in utilizing OWs greater than 120 minutes (and usually even 60), implying there are no additional predictive patterns to be found in the near past. On the contrary, the results from [1, 5, 7, 8, 15] show an accuracy decrease of only 1-2.5% when moving from a 60-min gap window to a 120-min one, which may imply that earlier observations may have a just a little less correlation to the target. Thus, there may be additional predictive patterns, which are not used in the right way by the existing methods.

2.6 Original Contributions

The main contribution of this paper is the In-Window Segmentation ($InWiSe$) method, which aims to utilize the OW local predictive patterns. The method, presented in Section 3.2 and Fig 1b, differs from previous methods by the following: (i) it extracts hidden local features by splitting the OW into multiple sub-windows, which improves the model predictive performance; (ii) it is flexible in terms of OW definition - if a complete sub-window set is not valid for use at the prediction time, a single OW option is used instead.

As mentioned in Section 2.3, a step towards multiple OW utilization was taken in [16] by combining weighted predicted posteriors of four OWs, each making an independent prediction with a distinct gap. Their approach is different from ours mainly in that we let the classifier learn the association between cross-window features, which is not possible in a weighted posterior decision. Another very recent work (DC-Prophet) [20], published while writing this paper, combines features from consecutive time series intervals (lags) to make early predictions of server failures. Their approach is similar to ours, but it has not been applied to the clinical episode prediction task, neither it has dealt with invalid lags.

A further contribution of our work is the evaluation of both our method and earlier episode prediction methods in a *cross-dataset* setting, in addition to the in-dataset cross-validation. Finally, our experiments are extended by a new evaluation approach, which excludes the clinically unnecessary in-episode predictions.

3 Methodology

This section starts with the problem definition, continues with introducing InWiSe, and concludes with the description of the data compilation, feature extraction, and classification methods used in this study.

3.1 Problem Definition and Prediction Modes

This study explores the problem of predicting a patient condition (hypotensive or normotensive) within a 60-min gap. Following the work in [1, 8] we define HE as a 30-min target window in which at least 90% of its MAP values are below 60 mmHg. Any valid target (see the validity definition in Section 4.2) not meeting this criterion is labeled as normotensive. At the prediction time, each OW (or sub-window set) is labeled with respect to its corresponding target.

Considering the implementation of the proposed method in the clinical setting, we distinguish between two alternative prediction modes: (i) *all-time prediction*, where the assumption (found in previous papers) is that episode prediction is needed continuously throughout the stay of a patient in the ICU, regardless of the clinical condition at the prediction time; (ii) *exclusive prediction*, where episode prediction is needed only when the patient is not in a currently recognized HE (the last 30 minutes of the ICU stay are not an HE by definition).

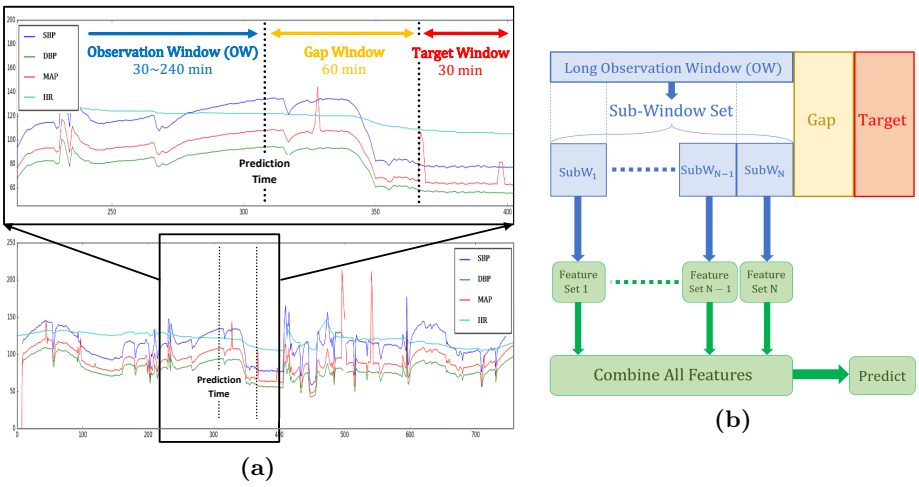

Fig. 1. (a) Basic Method: Traversing over a patient record with an impending HE is demonstrated by the observation, gap and target windows with respect to the prediction time. (b) InWiSe: a given OW is split into a sub-window set of size N, followed by a prediction that is based on the combined feature set of all N sub-windows ($SubW$s).

3.2 Splitting Windows with InWiSe

In our study, which was developed based on the observation-gap-target windows scheme demonstrated in Fig. 1a, we hypothesized that taking longer OWs, splitting them into several equally sized sub-windows, also called the *sub-window set*, and combining all their features together (see Figure 1b) would improve the predictive accuracy of the induced model verses using a smaller feature set of a single long OW. For example, a set of the mean MAP values from four, three, two and one hours before the same target window may be more informative for predicting the target label than the mean MAP value of a single 4-hour OW.

The InWiSe method does not use a classifier based on a combined set of features if one of the current OW sub-windows is invalid (see Section 4.2). In that case, the prediction is made by a simpler classification model using only the features extracted from the latest sub-window ($SubW_N$ in Figure 1b) unless that window is invalid. Consequently, InWiSe misses less prediction points than the single OW method (more about that in Section 4.4, in-dataset paragraph).

3.3 Feature Extraction

Three basic vital signs were used to derive two additional parameters for each record: Pulse Pressure calculated by $PP = SBP - DBP$, and Relative Cardiac Output calculated by $CO = HR \times PP$. Next, three group of features are extracted from each sub-window as detailed below.

Statistical features: mean, median, standard deviation (Std), variance, interquartile range (Iqr), skewness, kurtosis and linear regression slope are calculated for each of the 6 parameters (48 features). Missing values are ignored.

Wavelet features: similarly to in [5], multi-level discrete decomposition of each vital sign can be conducted with DB wavelets. The decomposition of a single time series (signal) X is denoted by $W_X = [a_n\ d_n\ d_{n-1}\ ...\ d_1]$, where n is the decomposition level (OW size depended), a_n is the signal approximation, and d_k is the detail signal of level k. The elements in W_X are then utilized as features by calculating the relative energy for each of them (a total of 24-42 features), as formulated by [8, 16]. Missing values are interpolated over the record time series.

Cross-correlation features: the cross correlation of two time series $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$ is defined by $\rho_{XY} = \frac{1}{n}\Sigma x_i y_i$ and calculated for each pair of vital signs (15 features overall).

The total amount of features extracted from a sub-window set is equal to the number of sub-windows N multiplied by the feature set size. For example, the feature dimension of the 4×60 configuration is $4 \times 98 = 392$.

3.4 Classification

Each instance in the training dataset is composed of a sub-window set feature vector and a class label which is positive or negative (the target is either hypotensive or normotensive, respectively). Before training a binary classifier, we both normalize the training dataset (to zero mean and unit standard deviation) and undersample it to overcome the imbalance obstacle (Section 4.3).

After discovering in our experiments that the Neural Networks and Random Forest methods are outperformed by the Extreme Gradient Boosting (XGBoost) [21] classifier (see Section 4.3), we chose the latter as the classifier for both methods. XGBoost is a scalable implementation of the Gradient Boosting ensemble method that affords some additional abilities, like feature sampling (in addition to instance sampling) for each tree in the ensemble, making it even more robust to feature dimensionality and helping to avoid overfitting. Moreover, considering this study minimum training dataset size of approximately 2.1k instances (Table 1) together with the maximal feature vector size of 392 features, the built-in feature selection capability of XGBoost is important.

Our classifier produces a posterior probability of the positive class, which may lead to an HE alert depending on the probability threshold determined from the Receiver Operating Characteristic (ROC) curve. The selection criterion for the optimal threshold can be like in [3, 8, 16], which is a popular indicator of classification accuracy: $Th_{selected} = \text{argmax}_{Th}\{sensitivity(Th) + specificity(Th)\}$.

4 Experiments

The experimental setup of this study is relatively branched due to the variety in prediction modes, methods and model configurations. We first perform an in-dataset evaluation for each prediction mode (all-time and exclusive) and for each method (single OW and InWiSe). Next, we assess a cross-dataset validation for each dataset pair. This section describes the datasets and their compiled OW and window-set statistics, followed by the experiments and analysis of results.

4.1 Data Description

Three databases of adult ICU stay records were prepared for this study: *Soroka* Medical Center in Beer Sheva (4,757 records), *Hadassah* Hospital, Ein-Karem, Jerusalem (8,366 records), and *MIMIC II* [10] (downloaded from [18, 19] and comprising 5,266 records). All time-series sampling rates are minute-by-minute (some second-by-second MIMIC II records were undersampled by taking each minute median). The common-shared vital signs among the three databases are HR, SBP, DBP, MAP, peripheral capillary oxygen saturation and respiration. As done by Lee and Mark [8, 16], we included only the HR, SBP, DBP and MAP vital signs in our data.

4.2 Data Compilation

As a pre-processing step, any outlier (out of the rage 10–200 for any vital sign) is considered as a 'missing value'. Consequently, henceforth mentioned missing values could have originally been outliers. When compiling OWs from each record we used the observation-gap-target windows scheme from Section 3.2 for the single OW method, as well as a first step for InWiSe (Figure 1b). The window sizes of the single OW method were 30, 60, 120 or 240 minutes, while the gap and target sizes were constant at 60 and 30 minutes, respectively. Furthermore, we followed Lee and Mark [8] who claimed that a prediction rate of every 30 minutes should reflect the performance of a real time system. Therefore, a 30-min sliding target window was traversed with no overlaps over each record and as many OWs as possible were compiled, depending on the prediction mode. As in [8], targets with more than 10% missing MAP values were excluded from this study, as their true labels are unknown. Turning to OW validity, to prevent from the classifier to learn outlier instances, more than 5% missing values for any vital sign made the window invalid and, consequently, excluded from our work as well.

Five configurations of window splitting were selected for this study: 60[m]→ 2x30[m], 120→2x60, 120→4x30, 240→2x120 and 240→4x60. For each configuration and for every record in the dataset, at the prediction time, we combine a

sub-window set ($[SubW_1, \ ... \ , SubW_N]$, Figure 1b) if all N sub-windows in the set are valid. The label of a complete sub-window set is the same as of its latest sub-window, which is labeled according to its corresponding target.

The data compilation step ended with the record and window label counts presented in Table 1. Noting that OW count ranges are window size dependent (30, 60, 120 and 240 minutes), and looking at the single OW method, we may see that the imbalance ratio for the all-time compilation is 1:20 to 1:40 in favor of normotensive (negative) windows, in comparison with the exclusive compilation (non in-episode prediction) which is double-imbalanced. As for the window set method, the bigger the set size (N) the less positive and negative examples are available. Again, we observe a worsening in imbalance which may reach a 1:100 ratio (Table 1 - MIMIC II, 4-window set, exclusive mode).

Table 1: OW and sub-window set label counts (in thousands) of each dataset, method and prediction mode. At the bottom are record label counts of each dataset

Method, Label	Mimic II		Soroka		Hadassah	
	All-Time	Exclusive	All-Time	Exclusive	All-Time	Exclusive
Single OW, HE	5.2–5.3	2.7	38–41	19–21	39–42	18–20
2-window Set, HE	4.6–4.7	1.6–2.1	37–39	12–18	36–40	9.5–15
4-window Set, HE	3.5–3.9	1.0–1.6	36–38	10–13	32–35	7.8–11
SIngle OW, Normotensive	184–193	181–191	1079–1174	1054–1153	828–899	805–897
2-window Set, Normotensive	157–167	152–164	1009–1109	951–1079	796–842	713–813
4-window Set, Normotensive	116–129	107–124	921–1003	854–960	672–764	621–707
Unstable Records	839		2,236		2,727	
Stable Records	1,612		2,098		2,823	

The reduction of sub-window sets availability with increasing N varies over datasets and is caused by differences in the amount of missing values (i.e., MIMIC II misses more values than Soroka). Moreover, the reason behind cross-dataset differences in terms of total OW count with respect to record count is the variance of ICU stay duration, which is higher in Soroka than in other datasets. Last, we note that using the exclusive mode results in a decrease of over 50% in the positive window count, probably because the average HE duration is much longer than 30 minutes (e.g. Hadassah average HE is 98 minutes long), increasing the time intervals where we do not make a prediction under this mode.

4.3 Experimental Setup

In-dataset Evaluation: For each dataset, mode and algorithm a 5-fold cross-validation (CV) was performed. To allow the classifier to successfully learn the imbalanced data, training folds were undersampled (5 times without replacement), and resulted in equal counts of stable and unstable records within each training fold (test folds were left unbalanced). Moreover, for each record all OWs or sub-window sets were either in training or test dataset to prevent record characteristics from leaking into the test dataset. In total, the classifier produced 25 outputs (5 folds $\times$ 5 samples) which were evaluated by five metrics: area under the ROC Curve (AUC), accuracy, sensitivity, specificity and precision. Furthermore, to compare between the two methods fairly, we optimize the

hyper-parameters of each method classifier: an inner 5-fold CV is utilized in each undersampled train fold of the earlier outer CV and the best chosen hyper-parameters found by a grid search are used to train the outer fold (Nested CV).

To choose the prediction model, three classifiers were evaluated using a 60-min OW (single OW method) and a 4x60-min set (InWiSe), on all datasets combined and in the all-time prediction mode (with optimization). The AUCs were (0.932, 0.936) for Random Forest, (0.937, 0.940) for Artificial Neural Networks (ANN) and (0.939, 0.943) for XGBoost, where each tuple represents a (single OW, sub-window set) pair. Since XGBoost significantly outperformed Random Forest with a p-value of $p = 0.03$ and ANN with $p = 0.08$, we chose XGBoost for this study (ANN was still used as a baseline from [8]).

The XGBoost classifier was grid-search optimized for each dataset or mode and for each OW size or sub-window set configuration C, where the best hyper-parameters were reproduced for all datasets, in most of the CV folds. The optimized hyper-parameters were: number of trees (500, **1000**, 1500), maximum tree depth (**3**, 5), learning rate (0.001, **0.01**, 0.1), instances sample rate (0.8, **0.4**) and features sample rate (0.9, **0.6**). The best choices are shown in bold.

Finally, each algorithm was tried with several OW sizes and sub-window sets Cs: four OW sizes for the single OW method and five Cs for InWiSe sub-window sets (see Section 4.2). As a result, a total of 54 in-dataset CVs were conducted (3 datasets $\times$ 2 modes $\times$ 9 window-set Cs and OW sizes).

Cross-dataset Validation: The model induced from each dataset was evaluated on other datasets using the all-time mode. XGBoost was trained on the entire dataset and tested on the two other datasets separately. The source dataset was undersampled only once, justified by a mostly very low variance of AUC (<0.1%) between undersamples, in each fold of the in-dataset CV. Both the OW size (single OW) and the window-set C (InWiSe) were chosen to optimize the AUC performance in the in-dataset evaluation: 120/240-min sized OW for the single OW method and 4x60-min sub-window set for InWiSe. The hyper-parameters of the classifier of each method were chosen by a majority vote over the folds in the in-dataset evaluation. A total of 18 experiments were performed (3 training datasets $\times$ 2 test datasets $\times$ 3 window-set Cs and OW sizes).

4.4 Analysis of Results

This subsection presents the results, followed by the feature importance analysis. The reader should note that all sub-window set results include test instances which were classified using the latest sub-window ($SubW_N$ in Fig.1b), if valid, in case that the sub-window set was invalid.

In-dataset: As a baseline, we reproduced the single OW method results of Lee and Mark [8] on MIMIC II with ANN. In Figure 2, we use MIMIC II to compare the single OW method with ANN, using two more methods: single OW with XGBoost (*single OW method*) and sub-window set best split with XGBoost as well. In comparison with the baseline, the AUC of the 4×60 sub-window set (XGBoost) was significantly higher than for the single OW method with ANN (60, 120 or 240-min OW size) with p-values 0.009 and 0.05 for all-time and exclusive modes, respectively.

From Figures 2 and 3, we first conclude that splitting a single OW into sub-windows is usually better than any single OW in terms of AUC; we note that, the advantage of OW splitting grows with an increase in the OW duration, which emphasizes the benefit from splitting a single OW that is longer than the longest OWs used by current methods (240→4x60 verses 120 minutes).

Turning to XGBoost, InWiSe outperformed the single OW with XGBoost in the all-time prediction mode, but only with a p-value of 0.13, while performing equally to the single OW method in the exclusive mode (AUC wise). However, while these all-time prediction trends are similar in the Soroka dataset where InWiSe is better with p=0.15 (Figure 3), in the Hadassah dataset InWiSe significantly outperforms the single OW method with a p-value of 0.03. In addition, in Table 2, that shows the in-dataset results in its diagonal, we see that, although not always statistically significant, the sub-window set method is better than alternatives in each dataset and in all metrics when evaluating in the all-time prediction mode. Note that our significance test comparisons were between the best results of each method rather than sub-window set verses its matching long window, as well as were calculated with a freedom degree of 4 (five folds).

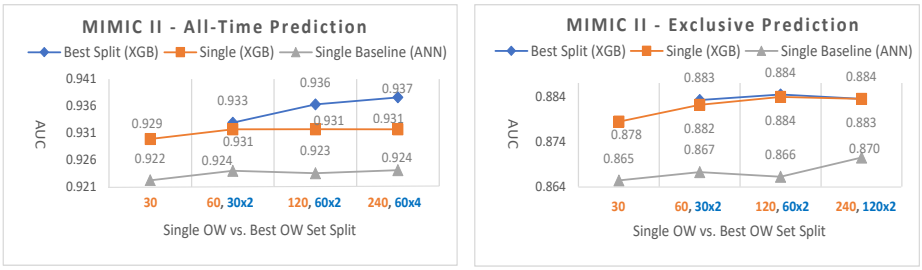

Fig. 2. The single OW method compared with its InWiSe best split, and with the ANN baseline on MIMIC II (all-time mode at the left and exclusive mode to the right).

As for the exclusive mode, the AUC was lower, as expected, since less positive OW instances are available, making the prediction task harder in general. The smaller *improvement* of InWiSe in comparison with a single OW is probably related to the large decrease in available positive sub-window sets relatively to the single OWs count, compared to the small decrease in the case of the all-time mode (Table 1). For example, Soroka HE labeled OWs count decreases by 5-7% for the all-time mode, but by 35-45% for the exclusive mode, between the 4-sub-window set and the single OW methods. Table 3 displays further metric results for the best OW size and sub-window set configuration of the single OW method and InWiSe, correspondingly. In contrast to the all-time in-dataset results in Table 2, here the AUC is still better with InWiSe (relatively to single OW), while other metrics domination is posterior probability threshold dependent.

Finally, we observed an average increase of 2.5% in valid prediction times when using InWiSe in comparison with a single OW in the size of N x $SubW_{size}$. This was mainly caused by the relaxed validation conditions in terms of missing values when splitting the windows, as well as by being able to use a single sub-

window instead of a sub-window set at the beginning of ICU stay when the available OWs are too short to be valid.

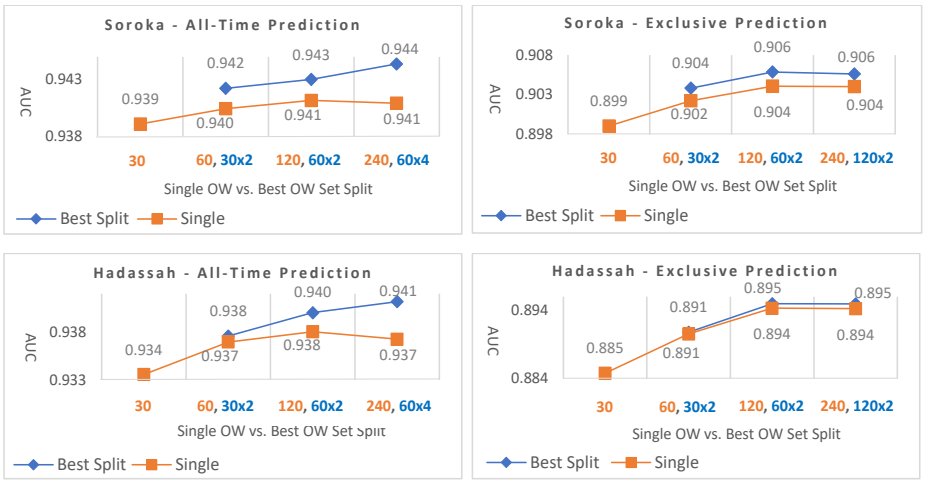

Fig. 3. In-dataset comparison between the single OW method and its InWiSe best split on the Hadassah and Soroka datasets (both prediction modes, XGBoost only).

Cross-datasets: The results of the cross-dataset experiments for the all-time prediction mode are shown in Table 2. First, one can observe the expected, but relatively small, drop in improvement when training with one dataset and testing with another (0.1–0.5% in AUC). Nevertheless, we see that InWiSe outperforms other methods in terms of the AUC metric, even when applying the model to a new dataset. However, similarly to the in-dataset exclusive mode case, the other metrics domination in the cross-dataset validation (all-time mode) is threshold dependent, but this time with dependence on the source dataset. For example, the Soroka dataset sensitivity of the single OW method is higher than the sub-window set one (0.897 vs. 0.868, respectively) in the case where the model was trained on MIMIC II, while the opposite is true when it was trained on Hasassah (0.920 vs. 0.940). The reason for these results is probably the difference between the optimal threshold values in the source and the target datasets.

Feature Importance: The goal of splitting OWs was to let the classifier learn feature correlations with the target window from each sub-window separately, as well as their cross-correlation with the target. Table 4 presents the top ten important features of XGBoost (frequent over the ensemble trees) for the best InWiSe configuration compared with its matching sub-window sized OW as well as the long single OW (in two prediction modes). The sub-window set columns are divided into their sub-windows, where $SubW_N$ is the sub-window ending at the prediction time and $SubW_1$ is the earliest in the set (see Figure 1b).

We first see that the MAP mean is clearly dominant in all cases, which make sense since MAP values are the ones that define an HE. Next, we observe that features from all three types (statistical, cross-correlation and wavelets) are top-

Table 2: In-dataset and cross-dataset methods comparison (InWiSe best configuration vs. single OW best size vs. window-set matching single OW size) using all-time prediction. Metrics from top to bottom: AUC, accuacy, sensitivity, specificty and precision

Target		Source Dataset (Train)								
		MIMIC II			Soroka			Hadassah		
		4x60	120	240	4x60	120	240	4x60	120	240
MIMIC II	AUC	**0.937±0.003**	0.931±0.006	0.931±0.007	**0.933**	0.929	0.932	**0.937**	0.933	0.934
	Acc.	**0.878±0.007**	0.864±0.013	0.866±0.011	0.846	0.828	0.860	0.898	0.903	0.902
	Sens.	**0.849±0.011**	0.847±0.025	0.843±0.023	0.869	0.873	0.849	0.829	0.807	0.803
	Spec.	**0.879±0.008**	0.865±0.013	0.866±0.011	0.846	0.826	0.861	0.900	0.906	0.905
	Prec.	**0.168±0.007**	0.150±0.011	0.151±0.016	0.136	0.124	0.146	0.187	0.194	0.192
Soroka	AUC	**0.939**	0.935	0.935	**0.944±0.003**	0.941±0.003	0.941±0.003	**0.942**	0.938	0.939
	Acc.	0.867	0.822	0.858	**0.870±0.007**	0.861±0.005	0.856±0.004	0.761	0.800	0.793
	Sens.	0.868	0.897	0.865	**0.875±0.014**	0.875±0.011	0.875±0.014	0.940	0.915	0.920
	Spec.	0.867	0.820	0.857	**0.869±0.008**	0.860±0.006	0.856±0.007	0.754	0.796	0.789
	Prec.	0.187	0.150	0.178	**0.191±0.015**	0.182±0.016	0.177±0.017	0.119	0.137	0.134
Hadassah	AUC	**0.936**	0.933	0.932	**0.938**	0.935	0.935	**0.941±0.002**	0.938±0.002	0.937±0.002
	Acc.	0.842	0.816	0.815	0.838	0.814	0.834	**0.870±0.006**	0.867±0.007	0.861±0.006
	Sens.	0.889	0.899	0.896	0.894	0.904	0.888	**0.871±0.006**	0.866±0.005	0.866±0.006
	Spec.	0.839	0.812	0.811	0.835	0.809	0.831	**0.870±0.007**	0.867±0.007	0.860±0.007
	Prec.	0.207	0.184	0.182	0.204	0.183	0.198	**0.241±0.011**	0.235±0.011	0.226±0.010

Table 3: In-dataset methods comparison (InWiSe best configuration vs. single OW best size) in the exclusive prediction mode

	MIMIC II		Soroka		Hadassah	
	2x60	120	2x60	120	2x60	120
AUC	**0.884±0.011**	0.884±0.006	**0.906±0.002**	0.904±0.002	**0.895±0.003**	0.894±0.003
Accuracy	0.790±0.020	**0.816±0.013**	0.787±0.011	**0.809±0.010**	0.783±0.009	**0.807±0.007**
Sensitivity	**0.826±0.021**	0.793±0.023	**0.875±0.009**	0.851±0.012	**0.855±0.009**	0.829±0.007
Specificity	0.789±0.021	**0.817±0.014**	0.785±0.010	**0.808±0.010**	0.782±0.009	**0.807±0.007**
Precision	0.054±0.006	**0.060±0.004**	0.069±0.004	**0.075±0.004**	0.082±0.003	**0.089±0.003**

Table 4: Top 10 features rank in terms of frequency over trees in the XGBoost ensemble

	All-Time Prediction						Exclusive Prediction			
	60-min	240-min	4x60-min				60-min	120-min	2x60-min	
			$SubW_1$	$SubW_2$	$SubW_3$	$SubW_4$			$SubW_1$	$SubW_2$
SBP Slope	10	10	–	–	–	9	–	10	–	9
SBP Skewness	–	–	–	–	–	–	9	–	–	–
DBP Slope	–	–	–	–	–	–	–	–	–	10
MAP Mean	1	1	5	–	6	1	1	1	4	1
MAP Std	6	8	–	–	–	–	8	9	–	–
MAP Median	2	3	7	–	–	2	3	3	7	5
MAP Iqr	9	–	–	–	–	–	10	–	–	–
MAP Skewness	8	7	–	–	–	–	6	6	–	8
MAP Slope	4	2	–	–	–	3	5	4	–	3
HR Slope	5	6	–	–	–	8	4	5	–	6
SBP Cross Correlation w/ DBP	7	5	–	–	–	10	7	7	–	–
PP Cross Correlation w/ RCO	3	4	–	–	–	4	2	2	–	2
MAP Wavelet Detail Level-3	–	–	–	–	–	–	–	8	–	–
HR Wavelet Approximation	–	9	–	–	–	–	–	–	–	–

ten-ranked, having the statistical ones (especially of MAP) used more frequently. Moreover, the two derived parameters, Pulse Pressure (PP) and Relative Cardiac Output (RCO), are proved to contribute particularly in their cross correlation with the target. Turning to sub-window sets, while $SubW_N$ has obviously more weight, the model repeatedly chooses to use the MAP mean and median from early sub-windows as well, and with a surprisingly high rank. These sub-window features are in favor of other features which are ranked higher in the single OWs (i.e., HR Slope and SBP cross-correlation with DBP). This evidence confirms our hypothesis that the classifier may be improved by using the local sub-window features instead of extracting the same features from a single long OW.

5 Conclusions and Future Work

The current study presented and explored InWiSe, an enhanced feature extraction algorithm for clinical episode prediction, where physiological features are extracted from a set of observation sub-windows instead of from a single OW. Our evaluation experiments have shown that the prediction performance may be improved by combining local sub-window features instead of extracting the same features from a single OW (of any size), observing an increased improvement when splitting longer OWs than in existing methods (i.e., 240-min OW).

The importance of sub-window features is confirmed by a feature importance analysis. Moreover, In all-time prediction mode, used by the recent works, we show an improvement in comparison with the single OW method over all three different datasets and for all measured metrics[5] (up to 1% in accuracy and specificity and up to 10% in precision, while maintaining the sensitivity equal or better). We particularly focus on the AUC metric that was improved by up to 0.6%. We also note the statistically significant improvement in AUC performance in the case of the Hadassah dataset.

In addition to the above, we successfully evaluated the methods in a cross-dataset fashion, showing that the AUC metric repeatedly favors the InWiSe method, even when testing the model on a new dataset. Furthermore, we explored a new prediction mode (exclusive) which may better reflect ICU needs.

With regard to InWiSe future improvement, better accuracy results may be achieved in the case of an invalid sub-window set, especially in the exclusive prediction mode. From the dataset aspect, any future analysis should use the recently published MIMIC III dataset mentioned in Section 2.6. Additionally, applying an existing model on a new dataset should be further investigated in terms of determining a dataset-specific optimal classification threshold.

References

1. Ghosh, S., et al.: Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure. IEEE J. Biomed. Health Inform., **20**(5), 1416-1426 (2016)

[5] All improvement percentages are in terms of a ratio between the two measures

2. Sebat, F., et al.: Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years. Crit. Care Med., **35**(11), 2568–2575 (2007)
3. Cao, H., et al.: Predicting ICU hemodynamic instability using continuous multiparameter trends. In: Eng. Med. Biol. Society (EMBS), pp. 3803–3806, IEEE (2008)
4. Moody, G. B., Lehman, L. W. H.: Predicting acute hypotensive episodes: The 10th annual physionet/computers in cardiology challenge. In: Comp. Card., pp. 541–544, IEEE (2009)
5. Forkan, A. R. M., et al.: 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)
6. Kamio, T., et al.: Use of Machine-Learning Approaches to Predict Clinical Deterioration in Critically Ill Patients: A Systematic Review. International Journal of Medical Research and Health Sciences **6**(6), 1-7 (2017)
7. Eshelman, L. J., et al.: Development and evaluation of predictive alerts for hemodynamic instability in ICU patients. In: AMIA Annual Symposium Proceedings, 2008, pp. 379. American Medical Informatics Association (2008)
8. Lee, J., Mark, R. G.: An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. Biomed. Eng. Online **9**(1), 62 (2010)
9. Tarassenko, L., et al.: Integrated monitoring and analysis for early warning of patient deterioration. British journal of anaesthesia **97**(1), 64–68 (2006)
10. Saeed, M., et al.: Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access ICU database. Crit. Car. Med. **39**(5), 952 (2011)
11. Chen, X., et al.: Forecasting acute hypotensive episodes in intensive care patients based on a peripheral arterial blood pressure waveform. In: Comp. Cardio., 2009, pp. 545–548. IEEE (2009)
12. Saeed, M.: Temporal pattern recognition in multiparameter ICU data, Doctoral dissertation, Massachusetts Institute of Technology (2007)
13. Saeed, M., Mark, R.: A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. In: AMIA Annual Symposium Proceedings, pp. 679. American Med. Info. Assoc. (2006)
14. Rocha, T., et al.: Wavelet based time series forecast with application to acute hypotensive episodes prediction. In: Engineering in medicine and biology society (EMBC), pp. 2403–2406. IEEE (2010)
15. Ghosh, S., et al.: Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. Journal of Biomed. Info. **66**, 19–31 (2017)
16. Lee, J., Mark, R. G.: A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. In: Comp. Cardio., pp. 81–84. IEEE (2010)
17. Rocha, T., et al.: Prediction of acute hypotensive episodes by means of neural network multi-models. Comp. in Bio. and Med. **41**(10), 881–890 (2011)
18. Goldberger, A. L., et al.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation **101**(23), e215–e220 (2000)
19. The MIMIC II Waveform Database Matched Subset (Physionet Database), https://physionet.org/physiobank/database/mimic2wdb/matched/.
20. Lee, Y. L., et el.: DC-Prophet: Predicting Catastrophic Machine Failures in DataCenters. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 64–76. Springer, Cham (2017)
21. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: 22nd ACM SIGKDD International Conference, pp. 785–794. ACM (2016)