applied sciences

MDPI

Article

Feature Engineering for ICU Mortality Prediction Based on Hourly to Bi-Hourly Measurements

Ahmed Y. A. Amer [1,2], Julie Vranken [3,4], Femke Wouters [3,4], Dieter Mesotten [3,4], Pieter Vandervoort [3,4], Valerie Storms [3,4], Stijn Luca [5], Bart Vanrumste [1] and Jean-Marie Aerts [2,*]

1 KU Leuven, E-MEDIA, Department of Electrical Engineering (ESAT) STADIUS, (ESAT) TC, Campus Group T, 3000 Leuven, Belgium
2 KU Leuven, Measure, Model & Manage Bioresponses (M3-BIORES), Department of Biosystems, 3000 Leuven, Belgium
3 Faculty of Medicine and Life Sciences, Hasselt University, 3500 Hasselt, Belgium
4 Ziekenhuis Oost-Limburg, Department of Anesthesiology, Department of Cardiology and Department Future Health, 3600 Genk, Belgium
5 Department of Data Analysis and Mathematical Modelling, Ghent University, 9000 Ghent, Belgium
* Correspondence: jean-marie.aerts@kuleuven.be

Received: 19 July 2019; Accepted: 24 August 2019; Published: 27 August 2019



Abstract: Mortality prediction for intensive care unit (ICU) patients is a challenging problem that requires extracting discriminative and informative features. This study presents a proof of concept for exploring features that can provide clinical insight. Through a feature engineering approach, it is attempted to improve ICU mortality prediction in field conditions with low frequently measured data (i.e., hourly to bi-hourly). Features are explored by investigating the vital signs measurements of ICU patients, labelled with mortality or survival at discharge. The vital signs of interest in this study are heart and respiration rate, oxygen saturation and blood pressure. The latter comprises systolic, diastolic and mean arterial pressure. In the feature exploration process, it is aimed to extract simple and interpretable features that can provide clinical insight. For this purpose, a classifier is required that maximises the margin between the two classes (i.e., survival and mortality) with minimum tolerance to misclassification errors. Moreover, it preferably has to provide a linear decision surface in the original feature space without mapping to an unlimited dimensionality feature space. Therefore, a linear hard margin support vector machine (SVM) classifier is suggested. The extracted features are grouped in three categories: statistical, dynamic and physiological. Each category plays an important role in enhancing classification error performance. After extracting several features within the three categories, a manual feature fine-tuning is applied to consider only the most efficient features. The final classification, considering mortality as the positive class, resulted in an accuracy of 91.56%, sensitivity of 90.59%, precision of 86.52% and F_1-score of 88.50%. The obtained results show that the proposed feature engineering approach and the extracted features are valid to be considered and further enhanced for the mortality prediction purpose. Moreover, the proposed feature engineering approach moved the modelling methodology from black-box modelling to grey-box modelling in combination with the powerful classifier of SVMs.

Keywords: feature engineering; intensive care unit; mortality prediction; hard-margin support vector machines

1. Introduction

Intensive care unit (ICU) patients are admitted because of an acute critical illness or because of the high need for intensive continuous monitoring. In addition, critical ICU patients are prone to

rapid deterioration, resulting in a possibly fatal outcome when not monitored closely. Hence, the main challenge at the ICU is to reduce the morbidity of the admitted patients and prevent mortality which has a high likelihood due to severe illness [1]. Mortality prevention requires an intensive monitoring of vital signs, such as heart and respiration rate, oxygen saturation, non-invasive or arterial blood pressure, and so forth, that can capture clinical deterioration earlier and thus improve patient outcome. In the past, multiple scoring systems have been developed (e.g., Acute Physiology, Age, Chronic Health Evaluation II, Simplified Acute Physiology Score, Sequential Organ Failure Assessment) to provide insights and even predictions regarding ICU patient mortality [2]. However, these scoring systems are population-based and often use summarised nongranular data. This calls for the need for an in-depth investigation of vital signs and associated indicators preceding any deterioration using granular continuous data. This investigation can be handled by time-series analytics to understand the behaviour and interaction of different signals.

Most of the ICU mortality prediction studies focus on developing powerful mortality prediction models [3–16] in which the higher priority is to provide an accurate label or score about the admitted patients' status. One drawback of such an objective is paying less attention to features' simplicity and interpretability, which is the case with deep learning approaches [7–13]. The key approach in these studies is black-box modelling focusing mainly on predictive model error performance, regardless the interpretability of the features. Hence, the useful information that can be provided to the medical staff is strictly the prediction output. Moreover, a considerable number of relevant studies focus on investigating the continuously recorded vital signs of ICU patients in order to predict the mortality–risk of those patients [3–16]. A frequently used database in these studies is the medical information mart for intensive care (MIMIC) in its three releases (MIMIC, MIMIC II and III) with different versions [17,18]. These databases provide a diverse and very large population of ICU patients and contain high temporal resolution data including lab results, electronic documentation and bedside monitor trends and waveforms. In contrast, another approach that is used in investigating critically ill patients in the ICU is mechanistic modelling [19,20]. Mechanistic modelling is used to describe the system from mathematical and physical dynamics perspective. The main focus of mechanistic modelling is on the system dynamics, the interaction between the different variables and the way they interact from a system perspective taking into account biological and physiological laws [21]. A mechanistic modelling approach is used in investigating biological systems by developing mathematical models [22–24].

The main focus in this presented study is to engineer features that can provide clinical insight by which the medical staff is guided through the different parameters. However, prediction accuracy is used in this study to assess the relevancy of the extracted features to the mortality events. Moreover, the dataset in our study is a low frequently measured data (i.e., hourly to bi-hourly) as it is a daily-life dataset that is not generated for research purpose. Moreover, the set of variables, parameters and the investigated population here is limited compared to the ones provided by the MIMIC databases. In the light of the given approaches (Black-box predictive models and mechanistic models) and reviewed studies, our study stands between the two approaches (i.e., pure black-box modelling and mechanistic modelling), as the main focus of the study is to achieve an efficient and informative set of physiologically meaningful features (mechanistic aspect) by means of enhancing the predictive model error performance (black-box aspect) that could be representable for European ICU departments.

From an analytical perspective, the series of recordings for each vital sign is considered a time-series that is sampled by a specific sampling rate. During ICU monitoring, different vital signs are measured and recorded simultaneously, in which the simultaneity facilitates studying correlation, interaction and behaviour between and within the different vital signs. Moreover, the time-series of recorded vital signs enable extraction of different features (typically statistical and dynamic) within segmented time windows, showing the dynamic behaviour of the recorded sign.

Many features can be extracted within consecutive or overlapping time windows for different vital signs, either individually or in combination. This option provides a large number of dimensions that have to be evaluated and adjusted to inform the decision making of the algorithm, which requires

an exhaustive investigation. However, such an investigation including a large number of numerical features is not an easy task for medical experts. Due to the high dimensionality issue, it is required to conduct such an investigation via a computational algorithm. In order to cope with these challenges, a simple and powerful classifier is used to explore the features. Ideally, this classifier should handle the problem of classification intuitively with the optimal margin hypothesis [25] which maximises the separability between the different classes. Moreover, the classifier should be capable of dealing with high dimensional data efficiently.

The proposed classifier for this purpose is the linear hard margin support vector machine (SVM) classifier which represents the simplest version of the powerful SVMs. The reason for using SVMs that it is relying on the maximum margin hypothesis. For linear hard margin SVM, it restrictively works efficiently once the input features provide linearly separable data points. With this property, it is feasible to extract features that may have a medical interpretation or physiological ground as the classifier would deal with the features as they are presented in the input space. In other words, it is required to have an acceptable performance only if the data points in the presented feature space are linearly separable with minimum misclassification error [25,26]. This error intolerance (or minimum tolerance) ensures that the introduced features provide a clear separation between the different classes (i.e., mortality and survival). Moreover, utilising such a linear classifier controls the dimensionality of the solution as it would only find a solution in the introduced dimensions. In other words, using a more sophisticated classifier (e.g., Radial Basis Function (RBF) SVM) would find a solution in an uncontrolled dimensionality, for instance, RBF SVM reaches infinite dimensionality due to the characteristics of the Gaussian kernel [26].

In this study, the problem is presented as integration between time-series prediction and classification. This integration is obtained by extracting features from the time-series and considering the dynamic behaviour of the time-series to construct the input space of the model. On the other hand, the output of the model is represented by the labels mortality/survival. The prediction is obtained by predicting the state (label) after the final record (last moment at ICU) on average 1.5 days ahead. The final record is the record preceding the patient's death (mortality label) or transfer to a lower care ward (survival label).

The objective of this study is to present a proof of concept for exploring features that can provide clinical insight through a feature engineering approach in order to improve the ICU mortality prediction in field conditions with low frequently measured data. The feature engineering approach is based on the hypothesis that utilising the linear hard margin SVM would provide a controllable and interpretable feature extraction approach.

This paper is arranged as follows: After the introduction, the second section of materials and methods comprises data description and an introduction to linear hard margin support vector machines. The third section includes the feature engineering process and results. The fourth section includes the discussion and the final section gives the conclusion.

2. Materials and Methods

2.1. Data

Data used for testing and evaluating the features were collected at the hospital Ziekenhuis Oost-Limburg (Genk, Belgium) during the period of 2015–2017. In detail, data were collected from patients hospitalised at the ICU and coronary care unit who were at these wards for at least ten days. Data consisted of vital parameters which were recorded continuously by Philips Intellivue monitors (Philips Electronics Nederland B.V., Amsterdam, The Netherlands), that recorded continuously and was annotated on average hourly to bi-hourly by the nursing staff. The recorded data was extracted from the electronic medical record for a total of 447 different patients, three of them readmitted to the unit again, hence, in total 450 recorded admissions annotated with either mortality or survival by discharge. The age of the patients was 65 ($\pm$16) years old, 305 of patients were males and 142 were

females. The average duration of stay at the ICU is 20.96 days with a minimum of 10 days, maximum of 97 days, median of 30 days and IQR of 20–53 days of ICU stay. The vital parameter data consisted of the heart rate, the respiration rate, oxygen saturation, arterial blood pressure (ABP), non-invasive blood pressure (if ABP was not measured) and body temperature (not frequently). The patient population of the study has different reasons for ICU admission as shown in Figure 1. The local Ethical Committee was notified and approval was obtained (19/0023R).

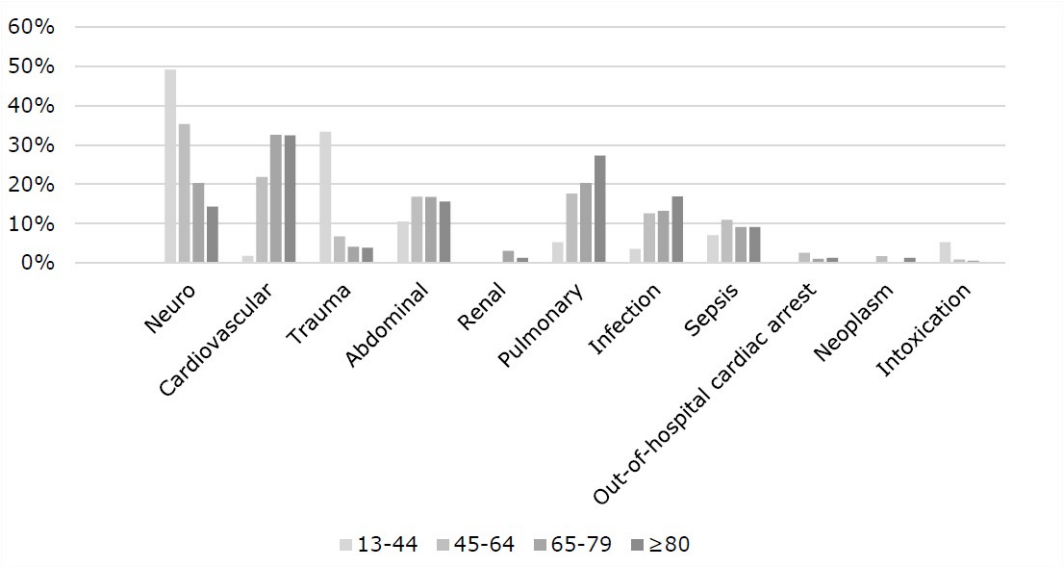


Figure 1. Distribution of the patient population and their reason for admission. The population was divided into age categories of 13–44, 45–64, 65–79 and >80 years of age.

2.2. Hard-Margin SVM

SVMs are originally presented as binary classifiers, that assign each data instance $\mathbf{x} \in \mathbb{R}^d$ to one of two classes described by a class label $y \in \{-1, 1\}$ based on the decision boundary that maximises the margin $2/||w||_2$ between the two classes as shown in Figure 2. The margin is determined by the distance between the decision boundary and the closest data point from each class [25–28].

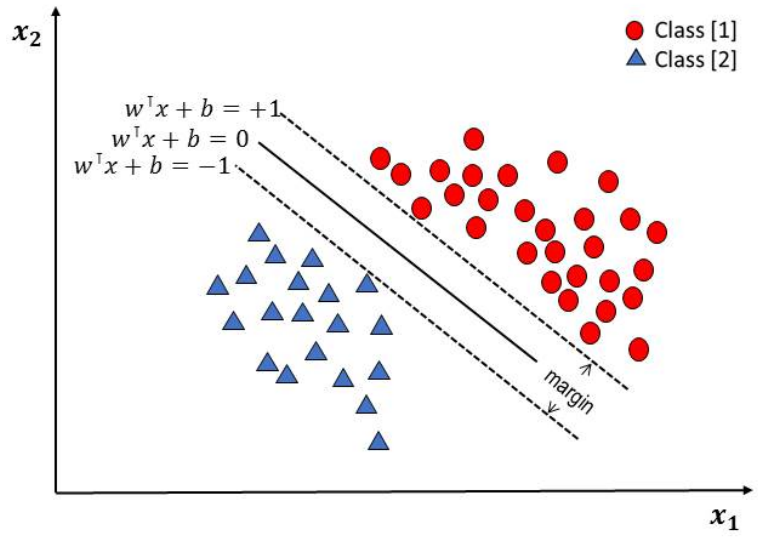


Figure 2. Schematic representation of a two-dimensional dataset consisting of two linearly separable classes. The dotted lines indicate the boundaries where the margin is maximised without tolerating any misclassifications (adapted from Reference [28]).

Generally, a feature map $\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathbb{R}^p$, where d is the number of input space dimensions and p is the number of feature space dimensions, is used to transform the geometric boundary between the two classes to a linear boundary $L : \mathbf{w}^\top \phi(\mathbf{x}) + b = 0$ in feature space, for some weight vector $\mathbf{w} \in \mathbb{R}^{p\times 1}$ and $b \in \mathbb{R}$. The class of each instance can then be found by $y = \text{sgn}\,(w^\top \phi(x) + b)$, where sgn refers to the sign function.

The estimation of the boundary L is performed based on a set of training examples $\mathbf{x}_i$ $(1 \leq i \leq N)$ with corresponding class labels $y_i \in \{-1, 1\}$, where N is the number of data points. An optimal boundary is found by maximising the margin that is defined as the smallest distances between L and any of the training instances. In particular, one is interested in constants $\mathbf{w}$ and b that minimise a *loss-function* [28]:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top \mathbf{w},$$

and are subject to:

$$y_i(\mathbf{w}^\top \phi(x_i) + b) \geq 1, \qquad i = 1, 2, ..., N.$$

By applying the lagrangian to the problem we get

$$\mathcal{L}(\mathbf{w}, b; \alpha) = \frac{1}{2}\|\mathbf{w}\|_2^2 - (\sum_{i=1}^{N} \alpha_i(y_i[\mathbf{w}^\top \phi(x_i) + b] - 1),$$

where $\alpha_i \geq 0$ are the Lagrangian multipliers for i^{th} data point. By solving the optimisation problem

$$\max_{\alpha} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b; \alpha),$$

the following optimisation conditions are obtained:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 &\longrightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i), \\
\frac{\partial \mathcal{L}}{\partial b} = 0 &\longrightarrow \sum_{i=1}^{N} \alpha_i y_i = 0, \\
\frac{\partial \mathcal{L}}{\partial \alpha} = 0 &\longrightarrow y_i(\mathbf{w}^\top \phi(x_i) - b) = 1,
\end{aligned}$$

The resulting classifier in both primal space and dual space are

$$f(\mathbf{x}) = \text{sgn}\,(\mathbf{w}^\top \phi(x) + b),$$

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{N} \alpha_i y_i \phi(x_i)^\top \boldsymbol{\phi}(\mathbf{x}) + b).$$

The dot product $\phi(x_i)^\top \boldsymbol{\phi}(\mathbf{x})$ is computationally expensive, hence, it is replaced with the *kernel* function $k(x_i, \mathbf{x})$, this replacement is known as the *kernel* trick. With the *kernel* trick, there is no need to execute the step of feature map as it is implicitly done by the *kernel* function. Hence, the dual space classifier with the *kernel* trick is

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{N} \alpha_i y_i k(x_i, \mathbf{x}) + b).$$

For practical reasons, we suggest to obtain the linear hard margin SVM from the standard SVM formula that tolerate misclassifcation errors [29]

$$\min_{\mathbf{w},b;\xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{N} \xi_i,$$

subject to:

$$y_i(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0, \quad i = 1, 2, ..., N.$$

where the constant C denotes the *penalty term* that is used to penalise misclassification through the slack variables ξ_i in the optimisation process. The linear hard margin SVM can be obtained via penalising the error extremely by giving C a very high value (e.g., 10^{10}). With this trick, we can get a solution with misclassified instances to be investigated through the feature engineering phase.

3. Feature-Engineering

The process of feature engineering is implemented in an interactive way between extracting new features and the classifier error performance as shown in Figure 3. This process is executed in three phases: feature-extraction, evaluation and feature fine-tuning. This process has a closed-loop nature as shown in Figure 3, since the three phases influence each other. The proposed three categories of features are statistical features, dynamic features, physiological features. The following sections describe the different feature engineering phases and the extracted features per category.

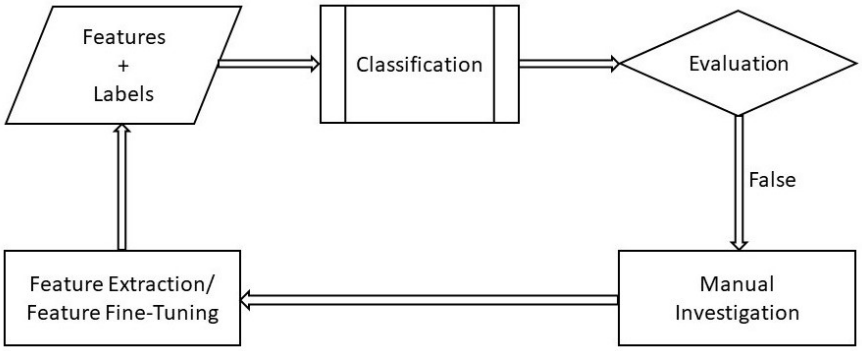


Figure 3. A flow chart illustrating the feature engineering methodology.

3.1. Evaluation

The engineered features are evaluated by feeding them into a linear hard-margin SVM classifier to predict mortality or survival of a subject. For this purpose, a leave-one-out procedure is used to produce a confusion matrix showing the true positives (TP), the true negatives (TN), the false positives (FP) and the false negatives (FN). The positive class is the mortality state and the negative class is the survival one. Using these numbers, different error performance metrics are calculated (i.e., sensitivity, precision, accuracy and F_1-score). Furthermore, we evaluate the features by looking at the effect on the number of true positives and true negatives when they are added to the model.

3.2. Feature Extraction

Firstly, all features are extracted within the last 84 observations which represent on average five days before the patient's discharge. The first 60 observations (3.5 days on average) out of 84 are considered for feature extraction to predict mortality/survival 24 observations ahead (1.5 days on average) at discharge (i.e., after observation 84). This period is determined after different test trials with different periods and is found to be the most efficient and informative period based on the classification performance. Moreover, this average period of 3.5 days agrees with the experience of clinical experts in the field. This agreement is based on the fact that there is no standard at the moment that refers to a minimum or maximum of observations to use, in order to provide the best of the care. As it is a

human/medical judgement which made based on a combination of patient-specific prognosis and trends, clinical expertise and experience and often corresponds to 3–4 days. The scheme of the feature extraction process is shown in Figure 4. Three categories of features are extracted, as described below.

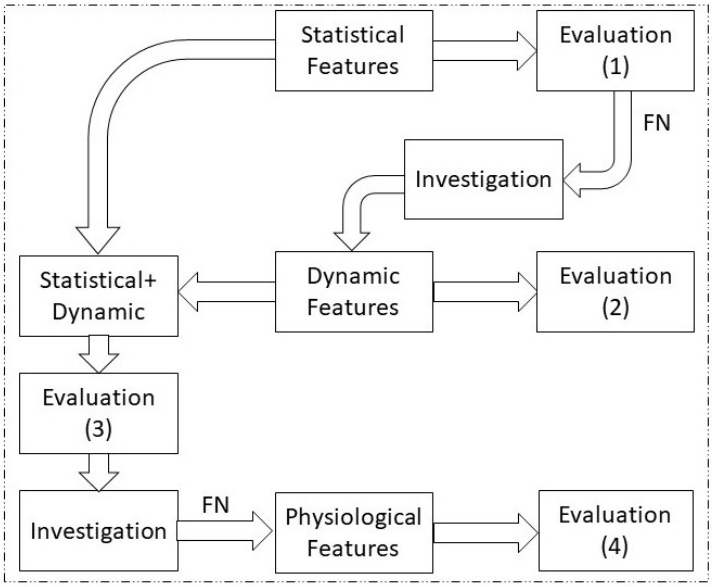


Figure 4. A flow chart illustrating the feature extraction process including the three feature categories (i.e., statistical, dynamic and physiological) and the sequence of the process marked by the evaluation steps. Also, in the process, the investigation is applied to the false negative patients only.

3.2.1. Statistical Features

The first category of features to be extracted is the set of statistical features which represent the basic characteristics of each time-series within segmented, non-overlapping time windows: *minimum*, *maximum*, *mean*, *median*, *standard deviation*, *variance*, and *energy*.

Statistical features are extracted within windows whose sizes are defined by the number of observations and not by a specific time period due to the nonuniform sampling rate (hourly to bi-hourly) as mentioned before. Extraction is based on the raw measurements of the vital signs and their first derivatives as well as the calculated standard early warning scores (EWS) of these measurements based on ZOL hospital standards. A weak point about statistical features is the static nature of these features as they do not reveal the dynamic behaviour of the time-series. Therefore, another category of features is required to be explored, namely dynamic features.

3.2.2. Dynamic Features

The extracted dynamic features are *Pearson correlation coefficients*, *crossing-the-mean count*, *outlier-occurrence count*, and *outlier indicator*. *Correlation coefficient* is computed between each pair of vital signs within each window. For this feature, it is necessary to be applied to the *z-score* of the vital signs. *Crossing-the-mean count* of a vital sign is determined by counting the number of times that the recorded vital sign crosses its mean value within each window. This feature indicates the abrupt changes in the vital sign from one observation to another. *Outlier-occurrence count* is computed by counting the number of outliers detected within each window. An outlier is detected by the statistical definition: any point outside the range $\mu \pm 3\sigma$ for a normally distributed variable is an outlier. For this feature, it is not expected to work with the vital sign of oxygen saturation (SpO_2) as it is negatively skewed, however, it will be tested as a feature to prove the concept. Finally, the *outlier indicator* is determined by the difference between the mean and the median of the records within each window.

3.2.3. Physiological Features

In order to enhance the classification performance, a manual investigation of the misclassified instances (based on the statistical and dynamic features) is required. The investigation is focusing on the false negative patients (i.e., deceased patients classified as survived) as the main objective is relevant to a reliable mortality prediction which is inversely proportional to the false negative count. This manual investigation is based on the measured physiological vital signs and uses physiological process knowledge resulting in physiological features. The different physiological features are described hereafter. By investigating the time-series of false negative patients, a consistent behaviour is noticed within the period of interest, in which the systolic blood pressure (SBP) approaches the diastolic blood pressure (DBP) as shown in Figure 5a.

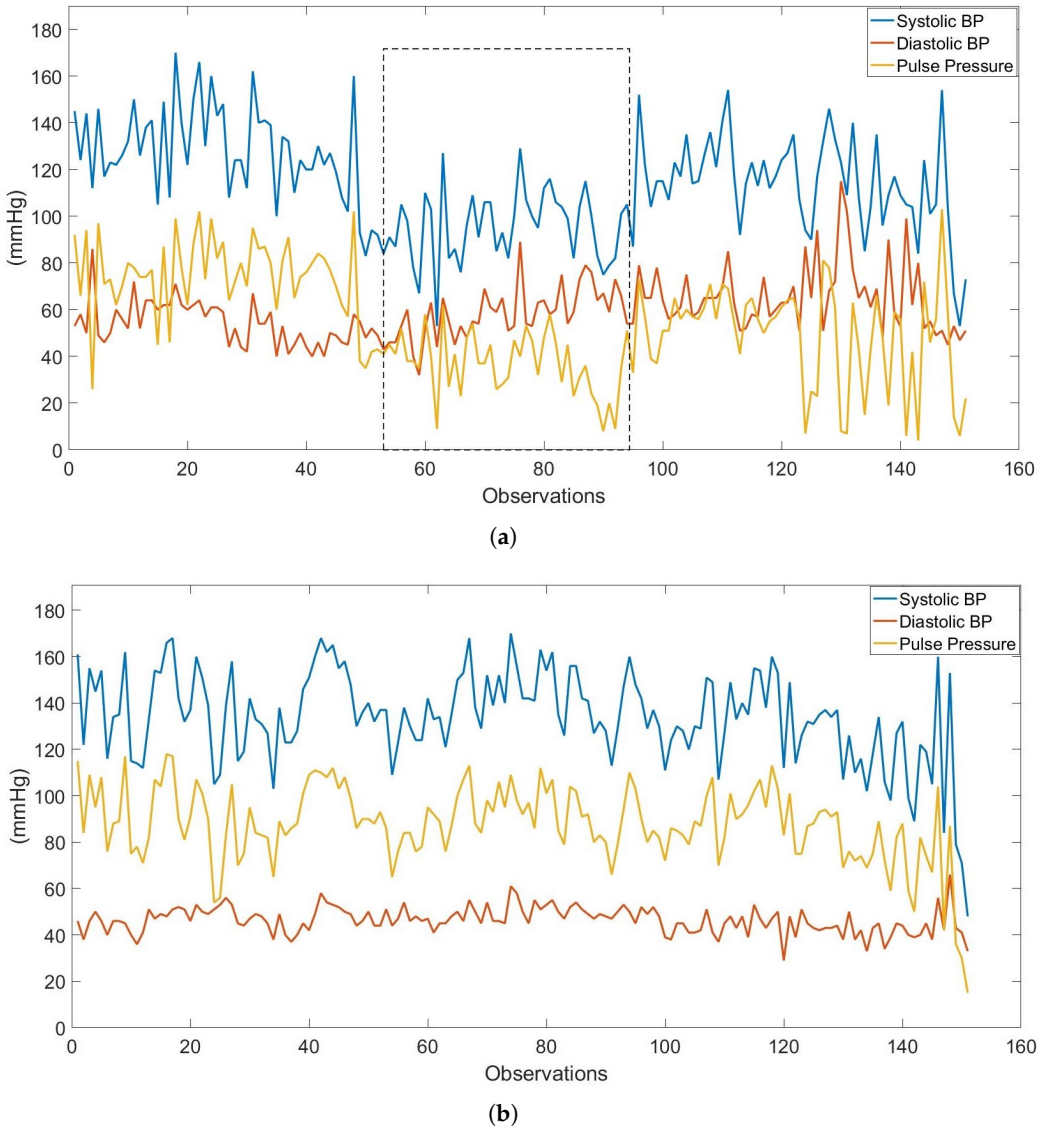
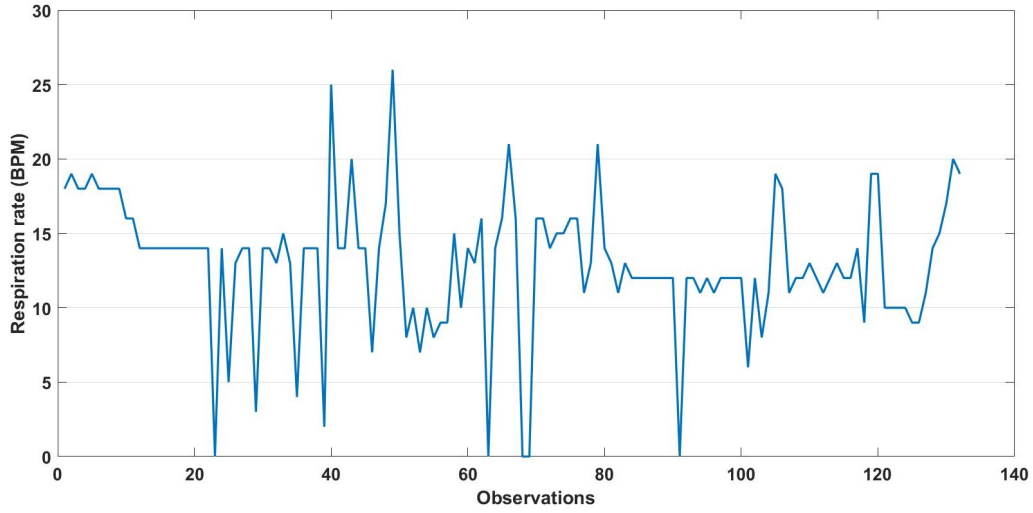


Figure 5. (**a**) Systolic blood pressure (BP), diastolic BP and pulse pressure (PP) of the last 150 observations (approximately the last nine days) of one false negative patient. The dashed window refers to the region where the systolic BP and diastolic BP measurements approach closely. (**b**) Systolic BP, diastolic BP and PP of the last 150 observations of another false negative patient. The mean value of the pulse pressure is 87.4 mmHg and median 88 mmHg.

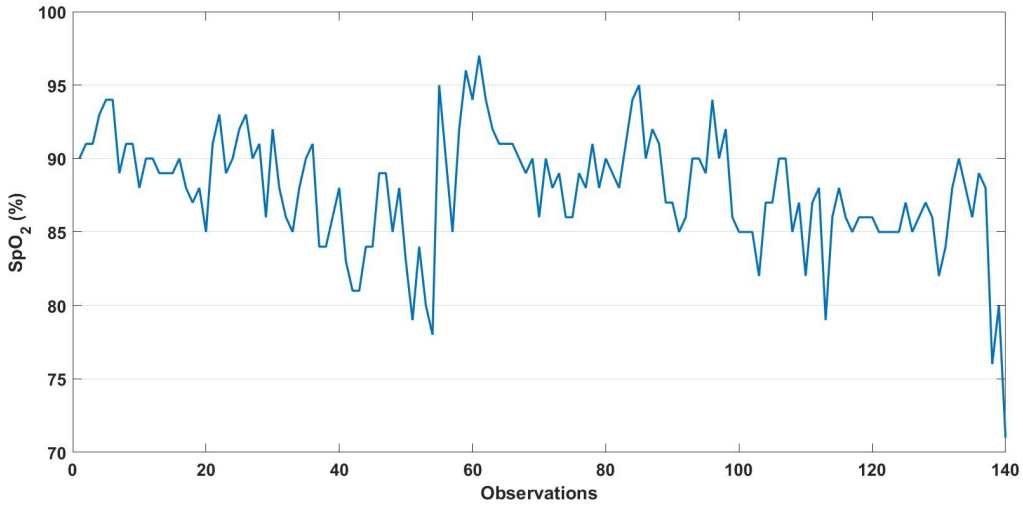
It is found that the difference (SBP-DBP) within certain measurement periods is smaller than 20 mmHg. A relevant observation that is noticed with other false negative patients is that this

difference is relatively high (i.e., greater than 60 mmHg) during certain measurement periods as shown in Figure 5b. This difference between SBP and DBP is also known as the pulse pressure (PP) and varies normally in a range between 40–60 mmHg [30,31]. As the PP is a linear combination between two vital signs, it can be considered as a new variable from which both statistical and dynamic features can be extracted. By reviewing medical literature focusing on PP and its effect on the mortality prediction (e.g., References [32,33]), our finding is partially consistent with their conclusion.

By further investigating the data, another behaviour is noticed with false negative patients, namely a frequent drop in respiration rate (RR) as shown in Figure 6a. Due to this behaviour, a new feature is proposed to represent this drop and the count of its occurrence. This feature is defined as the number of times the RR drops below a specific threshold within each window and is further referred as *low-RR count*. For this feature, two parameters are selected: the threshold and the window size. Both of them are searched exhaustively by maximising the classification performance by considering the new feature. The best-found combination is a threshold of 5 bpm and a window size of 60 observations.



(**a**)



(**b**)

Figure 6. (**a**) Respiration rate of a deceased patient with an obvious drop at specific observations below the normal range (12–20 BPM) (**b**) Oxygen saturation (SpO_2) of another deceased patient that drops frequently (minimum 78% during the stay).

Another observation in some false negative patients' vital signs is a physiological feature related to a frequent drop of oxygen saturation SpO_2 as shown in Figure 6b. Similar to *low-RR count*, this feature is defined as the number of times the SpO_2 drops below a specific threshold within each window. Moreover, the threshold and window size combination affects the influence of the feature on the performance. The best-found combination is a threshold of 77% and a window size of 60. This feature is further referred to *low-*SpO_2 *count*.

Both, *low-*SpO_2 *count* and *low-RR count* created only an added value to the classification performance after the fine-tuning step.

Finally, a physiological feature that is imported directly from the patients' medical record is their positive and negative diagnosis with cardiovascular diseases (*CVD*). By considering this feature exclusively in the input space, no single positive class is recognised. However, by adding this feature to the optimal combination of features, a remarkable enhancement is achieved as will be discussed later.

3.3. Feature Fine-Tuning

After defining three different categories of features, it is necessary to fine-tune the proposed features in order to obtain the most efficient combination and representation of them. As will be shown in Section 4, the error performance can drop after combining features from different categories. One interpretation of this drop is that some features are strictly efficient for a group of patients and confusing for the rest. In order to limit this effect a fine-tuning step is performed.

The feature fine-tuning phase is based on the selection of vital signs instead of the selection of dimensions which is in contrast with existing automatic and conventional feature-selection techniques. Indeed, the rows of the input matrix of our data correspond with the different subjects in the study and contain the different features calculated on multiple windows (e.g., the statistical feature of mean is extracted from m vital signs within n time-windows resulting in mn columns for each subject). Conventional feature-selection techniques select the columns of the matrix that are most representative for the study. However, in this way feature values within a specific time-window can be excluded leading to features that are hard to interpret. For this reason, we propose a backward selection approach where a feature (corresponding to multiple columns in the input matrix) can be excluded from the set of features. Moreover, prior knowledge is used in order to reduce the randomness in the selection process of the features. For instance, we will exclude the statistical and dynamic features of the HR guided by the prior knowledge that the heart is a main actuator in the control system of a human body that responds to different excitations (e.g., medication), not only critical events [34]. The effect on the performance score of this selection will be discussed in Section 4.

The procedure of feature fine-tuning that we propose in this work starts with exploring whether statistical and dynamic features are providing high performance when extracted from all vital signs or strictly from a subset of these vital signs. Moreover, we assess the effect on the classification performance of using aggregate features which are calculated on a group of vital signs together rather than on individual vital signs. Furthermore, feature values can be presented as either real or absolute. This procedure is applied exhaustively to the statistical and dynamic features and is assessed by the error performance. The resulting fine-tuning (FT) steps are as follows:

1. FT1: For HR extracted features, it is found that excluding both statistical and dynamic features enhances the error performance.
2. FT2: The *correlation coefficients* feature is found more efficient when presented in both real and absolute values.
3. FT3:*Outlier-occurrence count*, is found most efficient when applied to SBP, MAP, RR and PP excluding DBP and SpO_2. Moreover, the *outlier-occurrence count* is found more efficient when presented in an aggregate form instead of individually except for the vital sign SBP.

4. FT4: The *correlation coefficients* feature is providing the best performance when computed only between HR and SBP. Together with considering the features *low-*SpO_2 *count* and *low-RR count* the classification performance is improved.
5. FT5: *crossing-the-mean count* is found more efficient when applied only to SBP and RR and represented in the aggregate form.
6. FT6: The dynamic feature of *outlier indicator* is more efficient when applied only to SBP and DBP.
7. FT7: Ultimately, considering the physiological feature of CVD enhanced the performance.

4. Results

The obtained results based on the previously mentioned evaluation metrics for each category and for each fine-tuning step are explained below.

Starting with the statistical features, the resulting classification output is 83 TP's, 148 TN's, 87 FN's and 132 FP's. This result is fixed over the different test trials score-wise and patient-wise. In other words, the correctly classified patients are fixed over the different test trials because of using the linear hard margin SVM.

For dynamic features, the resulting classification output considering only the dynamic features is 32 TP's, 247 TN's, 138 FN and 33 FP's. Despite the remarkable reduction in the number of TP's, 18 new TP's are recognised by the dynamic features that are not recognised by the statistical features, in addition to 116 new TN's. This result is again fixed over the different test trials score-wise and patient-wise. With both statistical and dynamic features, the classifier performance is improved slightly compared to only statistical features with 2.8% increment in the accuracy. As the resulting classification output after combining both categories is 85 TP's, 159 TN's, 85 FN and 121 FP's. Despite the weak performance at this stage, the correctly classified instances are fixed with each test trial. This means that extracted features at this level are able to discriminate clearly between the correctly classified patients.

For physiological features, namely the PP, the resulting classification output with exclusively the extracted statistical and dynamic features of PP is 45 TP's, 222 TN's, 125 FN's and 58 FP's. It is important to note that the investigated FN's at the earlier stage are correctly classified by the PP extracted features. However, adding the PP extracted features to both statistical and dynamic features provided the following results: 83 TP's, 118 TN's, 87 FN's and 162 FP's. The classification output of the different feature-categories combinations are shown in Figure 7a. Moreover, feature extraction results are combined and depicted in Table 1.

Table 1. Feature Extraction results.

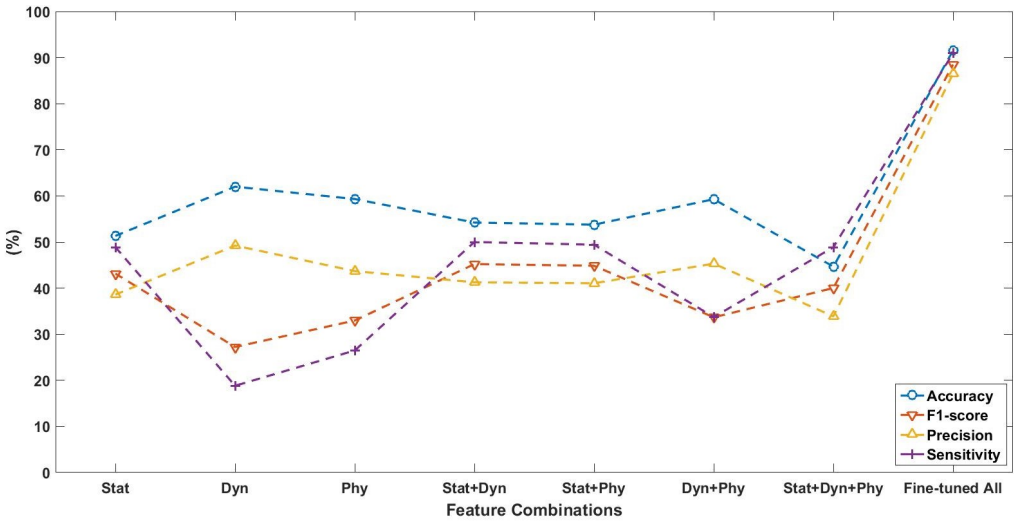
	Results							
Feature Combination	**TP**	**TN**	**FN**	**FP**	**Sensitivity (%)**	**Precision (%)**	**F1-Score**	**Accuracy (%)**
Statistical (Stat)	83	148	87	132	48.88	38.60	43.14	51.33
Dynamic (Dyn)	32	247	138	33	18.82	49.23	27.23	62.00
Stat+Dyn	85	159	85	121	50.00	41.26	45.21	54.22
Physiological (Phy)	45	222	125	58	26.47	43.69	32.97	59.33
Phy+Stat+Dyn	83	118	87	162	48.88	33.88	40.02	44.67

Before showing the results of the fine-tuning phase, we present the results of using the feature selection and ranking technique of automatic relevance determination (ARD) [28] based on backward selection method. The classification output of the ARD selected dimensions is 92 TP's, 218 TN's, 78 FN's and 62 FP's.

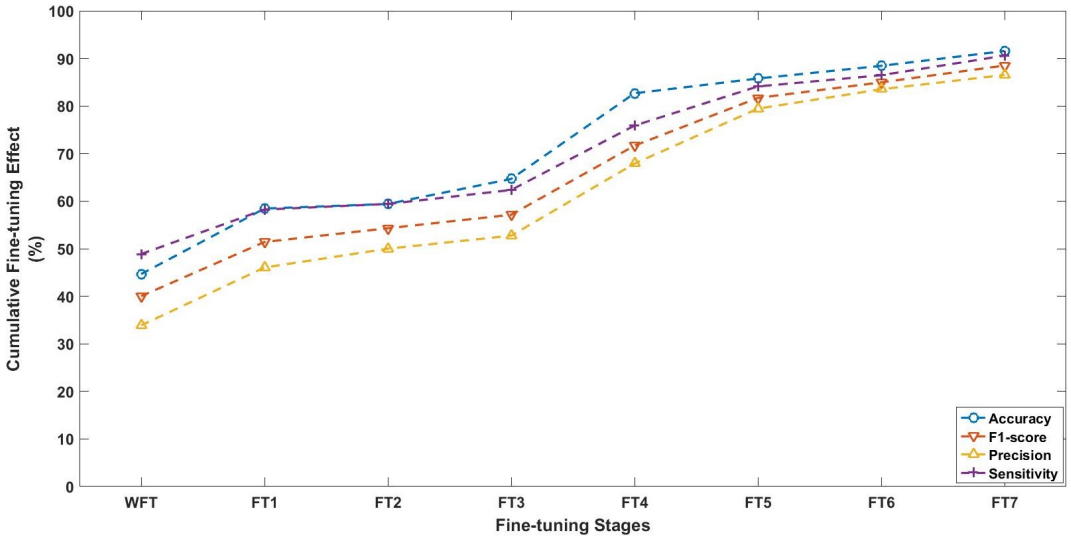
For the fine-tuning phase, the results are depicted in Table 2 and Figure 7b in a cumulative way.

Table 2. Feature Fine-tuning results.

	Results							
Cumulative Fine-Tuning Steps	**TP**	**TN**	**FN**	**FP**	**Sensitivity (%)**	**Precision (%)**	**F1-Score**	**Accuracy (%)**
ARD	92	218	78	62	54.12	59.74	56.80	68.89
FT1	99	164	71	116	58.23	46.04	51.42	58.44
+FT2	101	179	69	101	59.41	50.00	54.30	59.41
+FT3	106	185	64	95	62.35	52.74	57.14	64.67
+FT4	129	219	41	61	75.88	67.89	71.66	82.67
+FT5	143	243	27	37	84.11	79.44	81.70	85.78
+FT6	147	251	23	29	86.47	83.52	84.97	88.44
+FT7	154	256	16	24	90.59	86.52	88.50	91.56



(**a**)



(**b**)

Figure 7. (**a**) F_1-score, accuracy, sensitivity and precision of the classifier with all possible combinations of the three feature categories in addition to the fine-tuned combination. (**b**) The cumulative effect of the different fine-tuning stages on the classification accuracy, F_1-score, precision ,and sensitivity. WFT refers to 'without fine-tuning', FTx refers to the xth stage of fine-tuning as illustrated in the text.

5. Discussion

Many studies are using the area under receiver operating characteristics curve (AUC) as an evaluation metric. In this study, we prefer to use the confusion matrix for evaluation and direct quantification of error metrics of concern (e.g., sensitivity, precision). However, the calculated AUC for our optimised classifier is 0.91 for comparison purposes. This result, when compared to several recent studies is satisfactory. For instance, a recent study focusing on a special profile of ICU patients reported an AUC of 0.70 using a developed novel mortality prediction SOFA-RV [35]. Another study [12] that evaluates the Super ICU Learner Algorithm (SICULA) and its predictive power applied to MIMIC II database reported an AUC of 0.88 on average under specific conditions and 0.94 on average when applied to an external validation set with calibration. The study of Luo Y. et al. [11] reported an AUC of 0.848. Luo Y. et al. proposed an unsupervised feature learning algorithm that extracts features automatically from the clinical multivariate time-series. Luo Y. et al. applied their algorithm to the MIMIC-II [17] dataset with a prediction horizon extending to 30 days. The study in Reference [8] that developed a convolutional neural network (CNN) as a deep learning approach to predict mortality risk at ICU reported, as the highest performance, an AUC of 0.87, a precision of 0.7443 and a recall of 0.8188. The developed model used the variables of heart and respiration rate, systolic and diastolic blood pressure obtained from the MIMIC-III dataset [18]. Landon et al. [8] referred to the difficulties and limitations of using electronic medical report (EMR) data, similar to our dataset, for the purpose of mortality prediction at ICU. Nemati et al. in their study [36] of sepsis early prediction, which is a lead cause of morbidity and mortality of ICU patients, developed a machine learning model that reported an AUC of 0.83–0.85 for a prediction horizon of 12 down to 4 h prior to clinical recognition. Nemati et al. used an EMR data with high-resolution vital signs time-series obtained from the MIMIC-III dataset [18]. Two medical studies [32,33] reported an observed relevance between the low pulse pressure and mortality risk. Which is consistent with our finding of considering the pulse pressure as an independent variable from which both statistical and dynamic features can be extracted to inform mortality prediction. Moreover, the medical study in Reference [37] concludes the relevance between the widened (high) pulse pressure and the mortality risk for a special profile of critically ill patients. This conclusion as well is consistent with our finding, as we referred to the statistical and dynamic features of the pulse pressure which will indicate either abnormally high or low levels of pulse pressure. It is important to note that each study has different conditions, different objectives, different datasets, parameters and variables and predictive models.

At the feature extraction phase, the variation of results with different categories shows that a set of features can be efficient with a group of patients (i.e., correctly classified) but the same set of features can be inefficient or confusing to another group of patients (i.e., misclassified). For instance, statistical features classified correctly 83 TP's and 148 TN's, on the other hand, dynamic features classified correctly 32 TP's and 247 TN's. Considering the patient identity, it is found that dynamic features correctly classified 18 TP's and 116 TN's that the statistical features misclassified. The same observation is noticed with PP extracted features (45 TP's and 222 TN's) and those features extracted from both SBP and DBP together (72 TP's and 199 TN's). The difference in this situation is that PP is a result of a linear combination between SBP and DBP, however, PP extracted features correctly classified 14 TP's and 58 TN's that are misclassified by SBP/DBP extracted features. Hence, the influence of features should be evaluated on a subject-basis in addition to error metrics. Another observation is that the physiological features of *low-RR* and *low-SPO2 count* do not correctly classify any true positive patient despite their physiological basis when presented as the only input features. However, their contribution is significant when combined with the consistent set of features as shown at the feature fine-tuning phase. Therefore, excluding a feature has to be done after that it has been tested in combination with different groups of features especially if the extracted feature has a physiological basis.

At the fine-tuning phase, we have to note that this process is based on feature-vector-level not dimension-level as a single extracted feature may include multiple dimensions (e.g., the mean within each window for a specific vital sign). Which is in contrast with conventional feature selection

techniques that rely on selecting the most relevant dimensions regardless of the interpretation of the selected dimensions. The initial modification is excluding both statistical and dynamic features extracted from HR in order to enhance the performance. This modification is required as many of the cardiovascular patients in this study have common cardiac diseased behaviour, which confuses the classifier. Moreover, the heart acts as one of the main actuators in the human control system responding to different types of excitations. Hence, HR disturbances might not be sufficient to predict mortality, leading to a high false alarm rate. Ultimately, considering the cardiovascular patients specifically, HR statistical characteristics, as well as their HR dynamic features are both technically confusing to mortality prediction. Moreover, the enhancement of detecting more TP's by presenting some dynamic features in an aggregate form can be interpreted by the fact that the concurrence of vital signs deterioration is partially a sufficient mortality indicator but not a necessary one. In other words, total deterioration implies mortality but not vice versa. Introducing the *correlation coefficients* feature with absolute values in addition to real values provides an improvement. Both absolute and real values help the linear classifier to distinguish between the instances based on the correlation strength and correlation sign respectively. Restricting the *crossing-the-mean count* to SBP and RR caused an improvement. Thus, observation-to-observation variability of both vital signs even for a relatively low sampling rate (i.e., 0.5–1 sample/hour) is more informative than the other vital signs for resting patients such as ICU patients.

As the main objective of this study is to engineer feature that can provide clinical insight about mortality prediction, it is important to refer to the decision tree classifiers. As one of the decision trees advantages is model interpretability in terms of the input attributes. However, some shortages are present in decision trees in contrast with SVM's that supported the choice of the latter. These shortages are mainly the greedy nature of the algorithm, local optimisation, prone to overfitting and expensive computational cost compared to linear hard margin SVM in which there are no hyperparameters to optimise. Moreover, we based our study on the optimal margin hypothesis which is not provided by decision trees in contrast with SVM's. For comparison reasons, a decision tree analysis is applied to the final set of features. A CART algorithm decision tree (MATLAB 2017) is used with the following settings: the splitting criterion of *gdi*, minimum parent size of 368, minimum leaf size of 184, maximum splits of 450 and pruning based on classification error criterion. The classification output of the optimised decision tree is as follows: sensitivity of 41.2% precision of 42.42%, F1-score 41.80% and accuracy of 52.22%. It is obvious that the results are poor compared to the results of linear hard margin SVM. The poor performance is quite expected because of the conceptual differences between the two classification techniques (i.e., Decision trees and SVMs). It is possible that if the whole feature engineering process is designed based on the decision tree classifier properties, the results can be better.

Model development, feature extraction and fine-tuning are implemented on observation-basis instead of time-basis (hourly/daily). We hypothesise that observation-basis are more realistic as the events (observations) within a specific time period are more informative than time period regardless of the number of observations. Ideally, the number of observations is fixed along a specific period for all patients and uniformly distributed as well which is not the case with our dataset. However, for a proof of concept, we evaluate the classification performance based on extracting the same features on time-basis. Time-basis is implemented by considering the last 7 days before discharge, considering the first 5 days for feature extraction to predict mortality 2 days ahead. These periods are defined based on the observation-basis analysis. By extracting statistical, dynamic and physiological features without fine-tuning, the output classification performance is 88 TP's, 163 TN's, 82 FN's and 117 FP's. In comparison with the classification performance on observation-basis (83 TP's, 118 TN's, 87 FN's and 162 FP's) the error performance is higher.However, by following the same feature fine-tuning steps the final classification output (82 TP's, 160 TN's, 88 FN's and 120 FP's) is dropped compared to that obtained by an observation-based approach (154 TP's, 256 TN's, 16 FN's and 24 FP's). This drop can be interpreted by the fact that the fine-tuning phase is a manual crafting of the feature combination which is sensitive to the features setup (i.e., observation-basis or time-basis).

6. Conclusions

In this study, we proposed a proof of concept for a feature engineering approach to explore features that can provide clinical insight in order to enhance the mortality prediction of ICU patients using the machine learning algorithm of linear hard margin SVM. The optimal combination of features that provided the best classification performance comprises the following features:

1. Statistical features of the raw physiological variables, their first derivative of SBP, DBP, MAP, RR, SpO_2 and PP. Moreover, the statistical features extracted from the EWS of SBP, RR and SpO_2. A window size of 15 observations.
2. Real and absolute values of *correlation coefficients* between HR and SBP in a window size of 30 observations.
3. *Outlier-occurrence count* of SBP, MAP, RR and PP. represented in an aggregate form except for the SBP represented individually as well. A window size of 60 observations.
4. *crossing-the-mean count* of SBP and RR, it is presented in the aggregate form. A window size of 60 observations.
5. *Outlier indicator* of SBP and DBP. A window size of 60 observations.
6. *Low-SpO_2 count* less than 77% and *low-RR count* less than 5 BPM. A window size of 60 observations.

The proposed approach allows moving from black-box to grey-box modelling, starting from a powerful black-box technique such as SVMs. Moreover, in this case study, low frequently measured vital signs (hourly to bi-hourly) enabled us to extract efficient features for the purpose of relatively long term analysis.

From a feature engineering perspective, some features or variables are individually unable to distinguish between the two classes (i.e., mortality and survival). However, by combining such features in suitable feature combinations, their use becomes beneficial. Furthermore, combining different efficient features might cause a drop in performance. Therefore, a feature fine-tuning phase is essential in order to synthesise efficient feature-combination.

From the medical perspective, we can conclude that the heart rate as an individual variable can be confusing to predict the mortality. This conclusion is supported by improving the error performance by excluding the heart rate features. Moreover, we can recommend paying more attention to the pulse pressure explicitly, either high or low level, since both levels are found associated with the mortality of a group of patients. Watching the pulse pressure requires implicitly to consider the diastolic blood pressure which is excluded from the EWS standards. Finally, we conclude that different profiles of patients require a different set of features to handle the mortality prediction efficiently.

For future work, we propose to test the developed model with the extracted features along the stay of the ICU patients. In other words, we can scan the complete period of stay with the moving window of 60 observations for feature extraction to predict the mortality-risk 24 observations ahead. Despite the fact that along the stay the patients will be labelled as survival, the medical doctors may label any upcoming events with possible mortality-risk.

Author Contributions: Conceptualization, A.Y.A.A., P.V., S.L., B.V. and J.-M.A.; Data curation, J.V., F.W. and V.S.; Formal analysis, A.Y.A.A.; Investigation, A.Y.A.A.; Methodology, A.Y.A.A.; Resources, D.M. and P.V.; Software, A.Y.A.A.; Supervision, S.L., B.V. and J.-M.A.; Validation, A.Y.A.A.; Visualization, A.Y.A.A., J.V.; Writing—original draft, A.Y.A.A.; Writing—review and editing, A.Y.A.A., J.V., S.L., B.V. and J.-M.A.

Funding: This research is funded by a European Union Grant through wearIT4health project. The wearIT4health project is being carried out within the context of the Interreg V-A Euregio Meuse-Rhine programme, with EUR 2,3 million coming from the European Regional Development Fund (ERDF). With the investment of EU funds in Interreg projects, the European Union directly invests in economic development, innovation, territorial development, social inclusion and education in the Euregio Meuse-Rhine.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Braber, A.; van Zanten, A.R. Unravelling post-ICU mortality: Predictors and causes of death. *Eur. J. Anaesthesiol.* **2010**, *27*, 486–490. [CrossRef] [PubMed]
2. Goldhill, D.R.; McNarry, A.F.; Mandersloot, G.; McGinley, A. A physiologically-based early warning score for ward patients: The association between score and outcome. *Anaesthesia* **2005**, *60*, 547–553. [CrossRef] [PubMed]
3. Lokhandwala, S.; McCague, N.; Chahin, A.; Escobar, B.; Feng, M.; Ghassemi, M.M.; Stone, D.J.; Celi, L.A. One-year mortality after recovery from critical illness: A retrospective cohort study. *PLoS ONE* **2018**, *13*, e0197226. [CrossRef] [PubMed]
4. Celi, L.A.; Galvin, S.; Davidzon, G.; Lee, J.; Scott, D.; Mark, R. A database-driven decision support system: Customized mortality prediction. *J. Pers. Med.* **2012**, *2*, 138–148. [CrossRef] [PubMed]
5. Celi, L.A.; Tang, R.J.; Villaroel, M.; Davidzon, G.A.; Lester, W.T.; Chueh, H.C. A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury. *J. Healthc. Eng.* **2011**, *2*, 97–110. [CrossRef] [PubMed]
6. Johnson, A.E.W.; Mark, R.G. Real-time mortality prediction in the Intensive Care Unit. *AMIA Ann. Symp. Proc.* **2017**, *2017*, 994–1003.
7. Alves, T.; Laender, A.; Veloso, A.; Ziviani, N. Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation. *IEEE Int. Conf. Big Data* **2018**, 1328–1336. [CrossRef]
8. Landon, B.; Aditya, P.; Izzatbir, S.; Clayton, B. Real Time Mortality Risk Prediction: A Convolutional Neural Network Approach. *Int. Conf. Health Inf.* **2018**, 463–470. [CrossRef]
9. Zhu, Y.; Fan, X.; Wu, J.; Liu, X.; Shi, J.; Wang, C. Predicting ICU Mortality by Supervised Bidirectional LSTM Networks. In Proceedings of the IJCAI 2018 Joint Workshop on Artificial Intelligence in Health (AIH 2018), Stockholm, Sweden, 13–19 July 2018; pp. 49–60.
10. Johnson, A.E.; Pollard, T.J.; Mark, R.G. Reproducibility in critical care: A mortality prediction case study. *Mach. Learn. Healthc. Conf.* **2017**, *2017*, 361–376.
11. Luo, Y.; Xin, Y.; Joshi, R.; Celi, L.; Szolovits, P. Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
12. Pirracchio, R.; Petersen, M.L.; Carone, M.; Rigon, M.R.; Chevret, S.; van der Laan, M.J. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir. Med.* **2015** *3*, 42–52. [CrossRef]
13. Mayaud, L.; Lai, P.S.; Clifford, G.D.; Tarassenko, L.; Celi, L.A.; Annane, D. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. *Crit. Care Med.* **2013**, *4*, 954–962. [CrossRef]
14. Verplancke, T.; Van Looy, S.; Benoit, D.; Vansteelandt, S.; Depuydt, P.; De Turck, F.; Decruyenaere, J. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Med. Inform. Dec. Mak.* **2008**, *8*, 56–63. [CrossRef]
15. Kim, S.; Kim, W.; Park, R.W. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthc. Inform. Res.* **2011** *17*, 232–243. [CrossRef]
16. Vieira, S.M.; Mendonça, L.F.; Farinha, G.J.; Sousa, J.M. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Appl. Soft Comput.* **2013**, *13*, 3494–3504. [CrossRef]
17. Saeed, M.; Villarroel, M.; Reisner, A.T.; Clifford, G.; Lehman, L.W.; Moody, G.; Heldt, T.; Kyaw, T.H.; Moody, B.; Mark, R.G. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Crit. Care Med.* **2011**, *39*, 952–960. [CrossRef] [PubMed]
18. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Li-Wei, H.L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
19. Aerts, J.M.; Haddad, W.M.; An, G.; Vodovotz, Y. From data patterns to mechanistic models in acute critical illness. *J. Crit. Care* **2014**, *29*, 604–610. [CrossRef]
20. Young, P.C. *Recursive Estimation and Time-Series Analysis: An Introduction*; Springer Science and Business Media: Berlin, Germany, 2012.

21. Vodovotz, Y.; Csete, M.; Bartels, J.; Chang, S.; An, G. Translational systems biology of inflammation. *PLoS Comput. Biol.* **2008**, *4*, e1000014. [CrossRef]
22. Kumar, R.; Clermont, G.; Vodovotz, Y.; Chow, C.C. The dynamics of acute inflammation. *J. Theor. Biol.* **2004**, *230*, 145–155. [CrossRef]
23. Reynolds, A.; Rubin, J.; Clermont, G.; Day, J.; Vodovotz, Y.; Ermentrout, G.B. A reduced mathematical model of the acute inflammatory response: I. Derivation of model and analysis of anti-inflammation. *J. Theor. Biol.* **2006**, *242*, 220–236. [CrossRef]
24. Day, J.; Rubin, J.; Vodovotz, Y.; Chow, C.C.; Reynolds, A.; Clermont, G. A reduced mathematical model of the acute inflammatory response II. Capturing scenarios of repeated endotoxin administration. *J. Theor. Biol.* **2006**, *242*, 237–256. [CrossRef]
25. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
26. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
27. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
28. Suykens, J.A.K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle J. *Least Squares Support Vector Machines*; World Scientific Publishing Co.: Singapore, 2002.
29. Abu-Mostafa, Y.S.; Malik, M.-I.; Hsuan-Tien, L. *Learning from Data*; AMLBook: New York, NY, USA, 2012.
30. Homan, T.D.; Cichowski, E. *Physiology, Pulse Pressure*; StatPearls [Internet]; StatPearls Publishing: Treasure Island, FL, USA, 2018.
31. Stergiopulos, N.; Segers, P.; Westerhof, N. Use of pulse pressure method for estimating total arterial compliance in vivo. *Am. J. Physiol.-Heart Circ. Physiol.* **1999**, *276*, H424–H428. [CrossRef]
32. Yildiran, T.; Koc, M.; Bozkurt, A.; Sahin, D.Y.; Unal, I.; Acarturk, E. Low pulse pressure as a predictor of death in patients with mild to advanced heart failure. *Texas Heart Inst. J.* **2010**, *37*, 284–290.
33. Voors, A.A.; Petrie, C.J.; Petrie, M.C.; Charlesworth, A.; Hillege, H.L.; Zijlstra, F.; McMurray, J.J.; van Veldhuisen, D.J. Low pulse pressure is independently related to elevated natriuretic peptides and increased mortality in advanced chronic heart failure. *Eur. Heart J.* **2005**, *26*, 1759–1764. [CrossRef]
34. Grodins Fred, S. *Control Theory and Biological Systems*; Columbia University Press: New York, NY, USA, 1963.
35. Akin, S.; Caliskan, K.; Soliman, O.I.; Muslem, R.; Guven, G.; Van Thiel, R.J.; Struijs, A.; Gommers, D.; Zijlstra, F.; Bakker, J.; et al. A novel mortality risk score predicting intensive care mortality in cardiogenic shock patients treated with veno-arterial extracorporeal membrane oxygenation. *Eur. Heart J.* **2018**, *39*, 5690. [CrossRef]
36. Nemati, S.; Holder, A.; Razmi, F.; Stanley, M.D.; Clifford, G.D.; Buchman, T.G. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Crit. Care Med.* **2018**, *46*, 547–553. [CrossRef]
37. Al-Khalisy, H.; Nikiforov, I.; Jhajj, M.; Kodali, N.; Cheriyath, P. A widened pulse pressure: A potential valuable prognostic indicator of mortality in patients with sepsis. *J. Community Hosp. Intern. Med. Perspect.* **2015**, *5*, 29426. [CrossRef]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).