electronics

MDPI

Article

Two-Stage Monitoring of Patients in Intensive Care Unit for Sepsis Prediction Using Non-Overfitted Machine Learning Models

Vytautas Abromavičius, Darius Plonis *, Deividas Tarasevičius and Artūras Serackis

Department of Electronic Systems, Vilnius Gediminas Technical University, Naugarduko str. 41, 03227 Vilnius, Lithuania; vgtu@vgtu.lt or vytautas.abromavicius@vgtu.lt (V.A.); deividas.tarasevicius@vgtu.lt (D.T.); arturas.serackis@vgtu.lt (A.S.)

* Correspondence: darius.plonis@vgtu.lt

Received: 5 June 2020; Accepted: 9 July 2020; Published: 12 July 2020

Abstract: The presented research faces the problem of early detection of sepsis for patients in the Intensive Care Unit. The PhysioNet/Computing in Cardiology Challenge 2019 facilitated the development of automated, open-source algorithms for the early detection of sepsis from clinical data. A labeled clinical records dataset for training and verification of the algorithms was provided by the challenge organizers. However, a relatively small number of records with sepsis, supported by Sepsis-3 clinical criteria, led to highly unbalanced dataset (only 2% records with sepsis label). A high number of unbalanced data records is a great challenge for machine learning model training and is not suitable for training classical classifiers. To address these issues, a method taking into the account the amount of time the patients spent in the intensive care unit (ICU) was proposed. The proposed method uses two separate ensemble models, one trained on patient records under 56 h in the ICU, and another for patients who stayed longer than 56 h. A solution including feature selection and weighting based training on imbalanced data was proposed in this paper. In addition, several performance metrics were investigated. Results show, that for successful prediction, a particular model having few or more predictors based on the length of stay in the Intensive Care Unit should be applied.

Keywords: early detection; sepsis; evaluation metrics; machine learning; medical informatics; feature extraction; physionet challenge

1. Introduction

Sepsis is a syndrome of physiological, pathological, and biochemical abnormalities induced by infection [1]. The conservative estimates indicate that sepsis is a leading cause of mortality and critical illness worldwide [2,3]. World Health Organization concerned that sepsis continues to cause approximately six million deaths worldwide every year, most of which are preventable [4]. In their study, the Department of Health in Ireland reported that survival from sepsis-induced hypotension is over 75% if it is recognized promptly, but that every delay by an hour causes that figure to fall by over 7%, implying that the mortality increases by about 30%.

In this paper, we present our solution for the early detection of sepsis by joining the PhysioNet/Computing in Cardiology Challenge 2019 [5]. Here, a detailed explanation of the Challenge data, participant evaluation metrics, and primary results are provided, and therefore, we will not explain it in this paper. However, a few important findings we should share in this paper in order to better explain the motivation to construct our algorithm in a particular way.

According to the requirements of the Challenge, our open-source algorithm works on clinical data provided on a real-time basis by giving a positive or negative prediction of sepsis for every single hour.

The algorithm predicts sepsis development for the patient using a pre-trained mathematical model. Therefore, not only the appropriate model should be used but also the training should be performed in the right way.

Data used in the competition was collected from intensive care unit (ICU) patients in three separate hospital systems. However, data from two hospital systems only were publicly available for training (40,336 patients in total). Another set of records (24,819 patients in total), obtained from all three different hospital systems was hidden and used for official scoring only by challenges organizers. Such separation of the data prevented participants from over-fitting their models. Taking into account that the trained model may learn not only dependencies in the clinical records but also hospital system-related behavior, for our approach, we have tested different data selection strategies for training. Models trained on hospital system A data we tested on data from hospital system B and vice versa.

The most challenging issue in the available data records was a high number of unbalanced records. Only 2932 septic patients were included in the dataset, together with 37,404 non-septic patients. From the perspective of mathematical model training, the data balance is much worse. Since the sepsis prediction had to be made on an hourly basis, 6 h in advance to the onset time of sepsis, specified according to Sepsis-3 clinical criteria, a number of non-sepsis examples we also took from the septic patient early records. After such reorganization of training data, only 2% from 1,484,384 [1,424,171] events (16,933 from 752,946 [739,663] in set A and 10,557 from 73,438 [684,508] in set B) had to be classified as an early prediction of sepsis.

The imbalance of the data can be treated in different ways. Nemati et al. successfully used random subsampling to train deep cancer subtype classifier [6]. Vicar et al. used special cost function—Generalized Dice Loss [7]. Sweetly et al. created 54 datasets using the same sepsis data and different non-sepsis data records [8]. He et al. have applied a random subsampling to this Challenge data [9]. Although the rank of their solution was quite high in the Challenge, the model was highly overfitted on hospital systems A and B when comparing to the model performance on hidden hospital system C data. An interesting approach was proposed by Li et al., where they decided to divide data into three stages (1–9, 10–49 and above 50 h stay in ICU) [10].

Dealing with missing values is another decision to be taken and it also may have an influence to the selected model training and overall performance. Forward-fill method [8,11–15]. Singh at al. found in their study, that mean imputation model gave worst results [16]. Other authors successfully used mean calculation over whole dataset [15,17].

Our proposed algorithm was scored on a censored data set, dedicated for scoring and using utility function that rewards early predictions and penalizes late predictions as well as false alarms.

2. Materials and Methods

In this section, we address the challenges regarding the problem of early sepsis detection and propose a methodology to overcome them. A labeled clinical records dataset for training and verification of the algorithms was provided by the PhysioNet/Computing in Cardiology Challenge 2019 organizers [5].

2.1. The Data

Data contained records of 40,336 ICU patients with up to 40 clinical variables divided into two datasets, based on hospital systems A and B. For each patient, the data were recorded at every hour during the stay in ICU. The records were labeled (on an hourly basis) according to Sepsis-3 clinical criteria. A total of 1,407,716 h of data was collected and labeled. Data labels included vital signs, laboratory values, and demographic values of the patients. Eight vital signs were a heart rate (HR), pulse oximetry (O2sat), temperature (Temp), systolic blood pressure (SBP), mean arterial pressure (MAP), diastolic blood pressure (DBP), respiration rate (Resp) and end-tidal carbon dioxide (EtCO2). A total of 26 laboratory values were included in the dataset. Demographic values

include age, gender, hospital identifiers, the time between hospital and ICU admission (Hosp), and ICU length of stay (ICULOS). Data were labeled as positive 12 h before and 3 h after the onset time of sepsis. Positive labels of sepsis were found in 2932 of the 40,336 records, which is 7.27% of the data. Labels consisting of positive (sepsis) labels were found in 27,916 rows, which is only 1.98% of all data.

Investigation of the data showed large numbers of missing values. The percentage of missing rows of vital signs is shown in Figure 1. Missing values of vital signs make about 10% of the data, with the exception of Temp (66% missing data) and EtCO2 (100% and 92% missing data, for dataset A and dataset B, respectively). Therefore, EtCO2 was not used as a feature for the model. The percentage of missing rows of laboratory measurements is shown in Figure 2. Missing data of laboratory values makes from 78% to 100% for all values. We did not use laboratory values to develop our model.

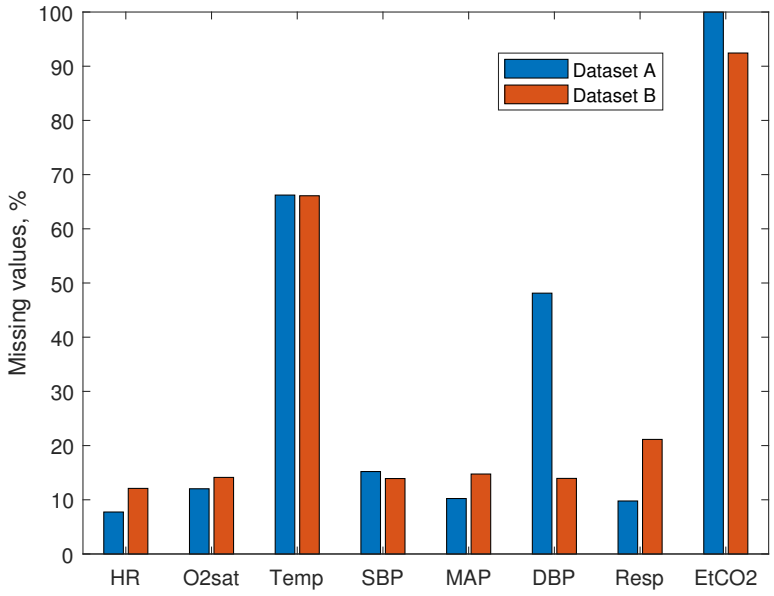

Figure 1. Missing vital values in the data.

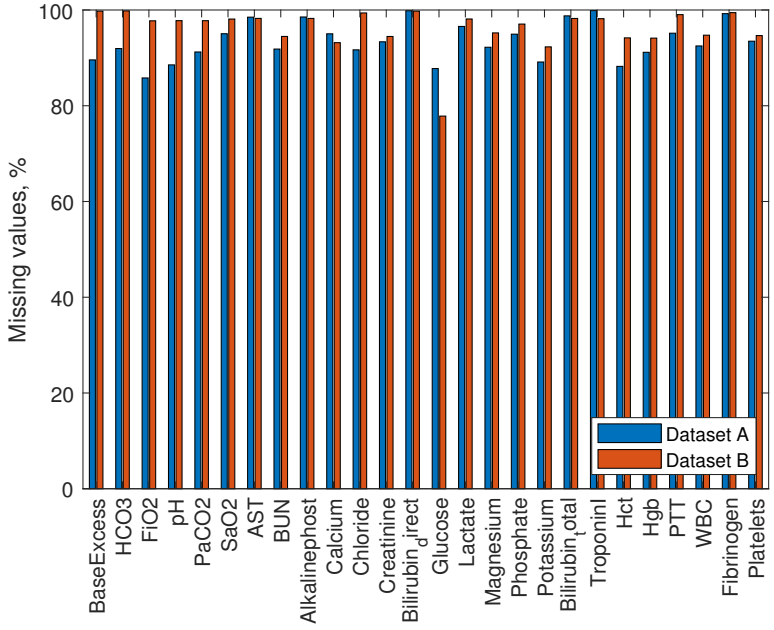

Figure 2. Missing laboratory values in the data.

Average values of vitals are shown in Table 1. Measured SBP, MAP and DBP values are higher in dataset B. Also, the measured HR is slightly lower in dataset B. Having two datasets collected in separate hospitals allowed us to develop models that are robust to measurement errors rising from the specificity of electronic medical record systems. Thus, the nature of data increases the difficulty of predicting sepsis. During the development of the model, we had to take into account the high unbalance of positive and negative cases, large amounts of missing values, and the fact that data was recorded using two different measurement systems.

Table 1. Average values of vitals and their standard error.

Measures	Dataset A	Dataset B
HR	84.6 ± 0.100	83.1 ± 0.106
O2sat	97.2 ± 0.016	97.1 ± 0.014
Temp	36.9 ± 0.004	36.8 ± 0.004
SBP	120.3 ± 0.115	126.4 ± 0.132
MAP	78.5 ± 0.076	86.7 ± 0.091
DBP	60.2 ± 0.070	66.6 ± 0.076
Resp	18.6 ± 0.026	18.5 ± 0.022
EtCO2	NaN	33.1 ± 0.072

2.2. Feature Extraction

A solution proposed in this paper to the early sepsis prediction problem employs information of the ICU length of stay, hospitalization time, age, and seven vital signs—HR, O2sat, Temp, SBP, MAP, DBP, and Resp. We did not use EtCO2 for feature extraction due to a large number of missing values.

We have calculated the mean, standard deviation, and the max-min difference for the vital sign data. We took those values from the whole duration of the record. Additionally, we have considered some other measures for our approach, such as kurtosis, entropy, and the standard error. However, after further analysis, we decided to discard these features. Kurtosis can only be calculated for four or more variables, not including the missing values. Additionally, kurtosis is not a representative statistic estimate for sample sizes less than 200 [18]. The entropy value is proportional to the sample size. In the problem we have investigated in this paper, the sample size changes each hour of the patient stay and can be reduced with missing values for some patients [19]. Therefore, in this case, the entropy just represents a number of samples used for its calculation. Thus, it is unlikely that entropy can carry useful information for the model training. The standard error is calculated by a division by the sample size, and it is inversely proportional to the sample size. Therefore, results can lead to a reduction of standard error for larger data sample sizes, which in its order increases unwanted load model training [20].

After our theoretical investigation, we have calculated 21 features for each hour. Missing values of the data were removed when calculating features. In some cases, features could not be calculated due to a small set of available data (e.g., during the first few hours of ICU stay, or due to a large number of missing values). In such cases, we set the value of the feature to '-1'. Finally, we have assembled a feature set of 24 features for model training: 21 calculated featured from vital signs and three demographic values (Hosp, age, ICULOS). Obtained data had different measurement units, measurement errors, and scales. Therefore, we have applied data standardization to have zero mean and standard deviation '1'. The sample mean and sample standard deviation were used the same for each patient, obtained from all sample data of both datasets containing 40,336 patients in total.

2.3. Data Balancing

The data from the experiment were strongly unbalanced, as discussed in Section 2.1. The balancing of the data can be performed using various oversampling or subsampling techniques that change

the data in the dataset before the model is trained. In our investigation, we followed the alternative approach by setting different weights for the individual data points according to their class.

During our investigation of possible data balancing approaches, we surveyed various sampling methods, also used in the Challenge by other authors. The challengers applied undersampling, subsampling methods and some oversampling. The immensity of the imbalance problem is very great; the ratio of the labels is 1:50. Additionally, clinical data is very contextual. Oversampling methods haven't shown good results in the Challenge. Undersampling methods possibly removed valuable information of non-septic patients. Most of the subsampling methods overtrained showed high Utility scores on known datasets and poor scores on hidden datasets. The variability of the clinical (ICU) data is very high. Non-septic cases have various other conditions and unknown prescribed medications. We think that adjustment of classification cost is a more robust and effective way to address this issue of data imbalance. The highest rank in the 2019 Challenge was received by a solution, where the classification cost was calculated using Utility function value differences between correct and incorrect classification [13]. In our solution, we used simple weighting and investigated the behavior of the model training with differently selected weighs.

We have weighed the positive and negative predictions in accordance with the duration before or after the onset time of sepsis. Positive values, with various weights, make only 1.98% of the data. For this reason, the investigated models were trained with uneven classification costs. We have addressed the data balance issue by utilizing and modifying the classification cost function:

$$loss = -\sum_{i=1}^{N} c_i \left| t_i - y_i \right|, \quad (1)$$

where N is the number of observations, t_i is the target output for observation i, y_i is the predicted value for observation i, and c_i is the classification cost for the observation i.

To investigate the influence of the selected weight on the model's behavior during training, we trained our models using different classification costs. A misclassified non-septic observations were weighted by 1 (c_1), and a misclassified septic observation weighted selecting different c_2 (1, 10, 20, 30, 100). Classification cost was selected based on the amount of "unbalance". As it was expected, since the positive values make only 1.98% of the data, the most effective value should be found when the classification cost c_2 is between 20 and 30. Models with c_2 parameter lower than 20 tend to predict most of the data as non-septic, while models with higher c_2 tend to predict all data as septic.

2.4. Model Training

In our investigation, we have used models based on Decision trees, naive Gaussian Bayes, Support Vector Machines (SVM), and Ensemble learners. All models were trained using stratified 5-fold cross validation. Decision trees based models were important in this stage of the problem solution. The Decision tree models give insights about the relevance of selected features. However, they tend to overfit the data [21]. Less over-fitting can be expected when using Ensemble learner models [22]. We have trained the ensemble learning-based models using hyperparameter optimization among Bag, GentleBoost, LogitBoost, AdaBoost, and RUSBoost methods. Hyperparameter tuning was performed for the number of decision tree splits, number of learners used in the model, learning rate, number of features in the ensemble. The number of decision tree splits search scaled in the range from 1 to 500. Many branches tend to overfit the data, while simpler trees can be more robust, which is especially important for the clinical data. The number of learners used in the search was from 10 to 100. A high number of learners can produce higher accuracy but can be time-consuming to train. We changed the learning rate in the ranged from 0.0001 to 1, and the number of features in ensemble hyperparameter tuning scaled from 1 to 24.

Gaussian naive Bayes based models are known for their simplicity, high bias, and low overfit. Typically good results using Naive Bayes are achieved using low variance data. These models are

not recommended for high variance data [23]. Models trained using SVM tends to overfit data less. However, they are not very successful in problems with a high number of missing values in data [24].

We have trained each model separately using features estimated on an hourly basis in random order. In order to avoid over-fitting and increase robustness, we have trained the models on records taken from a single hospital system (dataset A) and tested on records from another hospital (dataset B, hidden during the training). We trained models on the dataset A using 5-fold cross-validation. Using our selected approach, only half of the available data was used for training. However, as it was shown in the results of the challenge [5], proposed solutions to the problem performed well on known datasets, even if scoring was done on a hidden part of the same set, and performed marginally worse on new hospital system C, hidden from challenge contestants. Additionally, the advantage of this approach to train and evaluate models is supported by Biglarbeigi [25].

Based on the results of the trained models and insights into the data, we proposed a method that takes into account the amount of time the patients have already spent in the ICU. The first model used AdaBoost ensemble method with a decision tree to evaluate features of patients extracted during the first 56 h (11,146 sepsis labels and 654,866 non-sepsis labels; it makes approximate imbalance ratio of 1:59). The second model is applied for patients with ICULOS time greater than 56 (5990 sepsis labels and 118,213 non-sepsis labels; it makes an approximate imbalance ratio of 1:20). The second model was developed using the discriminant subspace-based ensemble method. The method generates decision trees using pseudorandomly selected feature components. Decisions of the trees are then combined by averaging the estimates. As this method is based on decision trees, it is fast to train and easy to interpret [26]. Both models were using features described in Section 2.2. Additionally, our proposed model was trained on both datasets, 75% of the data used for training. This allowed us to investigate how much model overtrains on known hospital systems.

2.5. Model Scoring

Models, proposed in this paper for Sepsis prediction, were evaluated using several different metrics. Traditional scoring metrics, such as the area under the receiver operating characteristics (AUROC), the area under the precision-recall curve (AUPRC), accuracy, F-measure, and Matthews correlation coefficient (MCC) were used. Additionally, investigated models were scored using a specific scoring function developed by the authors of the dataset, called Utility score. Performance of the investigated models was based on the Utility score metric. Additionally, using different scoring metrics allowed us a better comparison of investigated models. AUPRC is recommended for imbalanced data over the AUROC measure [27,28]. F-measure is a harmonic mean of precision and recall [29]. Lately, MCC measure was shown to be more advantageous over F-measure in the binary classification of imbalanced data [30].

The utility score metric was proposed by the authors of the dataset for the 2019 physionet Challenge [5]. This metric was recently proposed. The utility score metric reward algorithms that facilitate early sepsis detection and treatment. Additionally, it addresses the problem of infrequent events and sequential prediction tasks. It is designed to capture the clinical utility of early sepsis detection by weighting early and late predictions. Moreover, decision threshold metrics (AUROC and AUPRC) have problems evaluating unbalanced datasets. Additionally, other challengers evaluated their models using the Utility score, therefore, it is easier to compare the results. However, the experimental results showed that Utility score correlates with two traditional metrics F-measure and MCC.

Utility score—a specifically designed scoring function rewards algorithms for early predictions and penalizes them for late or missed predictions and false alarms. Scoring was conducted by predicting each hourly label for each patient. Each positive label had a defined score depending on correct prediction time to sepsis. Scoring function awarded models for correct prediction at most 12 h before and 3 h after the onset time of sepsis. Scoring function penalized models who predicted septic state 12 h before onset time of sepsis and slightly penalized models with false-positive predictions.

True negative predictions were not penalized or rewarded by the function. Best Utility score the most optimal model could achieve would be 1. Thus, a better model would have higher Utility score. A more detailed scoring description can be found in the original paper. The utility score was a reference metric. Using it, we evaluated the performance of our investigated models. AUROC, AUPRC, accuracy, F-measure, and MCC scores were used to gain insight into the models (e.g., if they correlate with any other parameter of the experiment, such as classification cost, feature reduction, model configuration, or Utility score).

Our investigated and proposed models were compared with five challengers who received the highest Utility score on hidden test C. Additionally, a result of three baseline models was generated. Baseline models were scored using all positive, all negative and random performance to clearly show the unbalance of the data and the difficulty of the challenge.

3. Results

As it was noted in Section 2.4, decision trees, naive Gaussian Bayes, SVM and ensemble learners were investigated in our experiment. Various parameters of the models were adjusted. Additionally, the effect of classification cost and feature reduction was investigated. The performance of developed models was evaluated using Utility score as a reference metric. To complement, several other metrics, such as AUROC, AUPRC, accuracy, F-measure, and MCC were calculated to compare investigated models. Models were trained using dataset A, and scored using dataset B. Results of the experiment are given in Table 2. Models, based on decision trees, are labeled from 1 to 14. Models, based on the Naive Bayes algorithm, are labeled from 21 to 25. SVM based models are labeled from 31 to 34. Models from 41 to 44 were using an optimizable ensemble method for searching the best model for the problem.

The random guess of the sepsis with an accuracy of 50% showed −0.529 Utility score and 0.125 AUROC. By labeling all cases as positive, the Utility score was reduced to −1.059 and AUROC was 0, F-measure—0.029. Labeling all cases as negative increased the Utility score to 0, AUROC was 0.5 and F-measure—0. AUROC was generally expected to be equal or above 0.5 when the dataset is balanced. In Table 2, the third row indicates the AUROC value of a random performance, having an accuracy of 50%). The AUROC value in such a case is 0.125 (for balanced datasets it would be 0.5). It is a direct insight into the importance of the problem.

Based on the results of the Physionet challenge 2019 on the hidden dataset, several of the highest Utility scores were from 0.017 to 0.193. The highest Utility score (0.193) was achieved by Hong et al. using deep recurrent reinforcement learners [31]. Murugesan et al. applied XGBoost algorithm and achieved 0.182 Utility score on hidden test set [32]. Our proposed method was developed based on the results of further described investigations and results. The proposed method predicted 3753 true positive, 707,606 true negative, 7027 false positive and 43,309 false negative observations of dataset B. The precision of the method was 34.82%, recall—7.98%. Our proposed method achieved 0.306 AUROC, 0.009 AUPROC, 0.934 accuracy, 0.129 F-measure, 0.142 MCC and 0.245 Utility scores. Additionally, this model was trained on both datasets (70% of the data used for training) achieved 0.276 Utility score.

Decision trees are fast to train and to evaluate. We started our investigation using these models. The baseline score of Model1, with default parameters, gave a Utility score of 0.01. Secondly, a feature reduction using principal component analysis (PCA) was applied. Six features to explain 95% variance was kept. Model2 gave Utility score of 0.004. For Model3, increasing feature set to 14 (out of 24) increased Utility score to 0.0124. Forth model used 14 features and a modified classification cost ratio of 1:10. Model4 obtained Utility score of 0.1236. Further increasing classification (Model5) ratio to 1:100 Utility score decreased to −0.296. Using all available features in the set (24 features) Utility score was slightly improved to −0.242, for Model6. Using 24 features and classification cost 1:10 obtained Utility score was 0.184 for Model7. For Model8 we modified classification cost to 1:20, obtained Utility score was 0.22. Further increasing classification cost to 1:30 (Model9) decreased Utility score to 0.216.

Model10, having a modified classification cost of 1:20, and a reduced feature set to 20 got Utility score decreased to 0.157. Next, we limited the tree split criterion to 50. Model11 achieved Utility score of 0.232. Higher Utility score was achieved by reducing split criterion to 4, it was 0.233 (Model12). Reducing split criterion to 2 for Model13 got a similar Utility score of 0.233. Model14, with a further reduced split criterion to 1, achieved the highest Utility score of 0.242. Only 1 tree branch was used for this model. A feature that was used for this model was ICULOS. Features used for Model12 and Model14 also included ICULOS, and also mean SBP and Resp.

Models labeled from 21 to 25 were based on the Naive Bayes algorithm. Model21 without feature reduction and using classification cost 1:20 achieved Utility score of 0.1334. For Model23, using PCA, the number of features was reduced to 14, achieved Utility score was 0.129. Further reducing the number of features to 6 (95% explained variance using PCA) improved Utility score to 0.150, as shown in Table 2 Model24 row. Adjusting the classification cost led to a reduced score—Model22 and Model25 used a reduced feature set and a modified classification cost of 1:10 and 1:30; they yielded Utility scores of 0.097 and 0.143, respectively. SVM models were computed using the Gaussian kernel function. Results of the SVM models are shown in Table 2, under Model31 to Model34. Model31, using a classification cost ratio of 1:20 and 24 features for training, achieved a 0.151 Utility score. Model32, using 6 features to explain 95% variance, achieved a 0.144 Utility score. Model33 and Model34, using classification costs 1:30 and 1:10, respectively, achieved 0.1294 and 0.1302 Utility scores.

Models labeled from 41 to 44 were trained using ensemble methods, searching between Bag, GentleBoost, LogitBoost, AdaBoost, and RUSBoost methods and other hyperparameters. The classification cost for all investigated models were set to 1:20. Model41 using a full feature set achieved Utility score of 0.082. Reducing the tree split criterion to 10 (Model42) gave an improved Utility score of—0.124. Ensemble model (Model43) using bagged decision trees, having 29 learners, 4 splits and using 24 features achieved a Utility score of 0.173, further reducing split criterion to 1 did not improve the Utility score—0.173. Using principal component analysis (95%) reduced the Utility score to 0.008 (Model44).

High AUROC, AUPRC and accuracy scores using decision tree models were achieved when classification cost was 1:1, for example, 0.492, 0.497, 0.491 AUROC score for Model1, Model2, Model3, respectively. Model12 and Model14 with high Utility scores gave low AUROC (0.313 and 0.309, respectively), AUPRC (0.009, both) and accuracy (0.931 and 0.937, respectively) scores.

High AUROC, AUPRC, and accuracy scores using ensemble learners were achieved using Model41: 0.413, 0.012, and 0.955, respectively. However, this ensemble model achieved low Utility (0.082). In the same manner, low AUROC (0.347), AUPRC (0.01), and accuracy (0.939) scores, and the highest Utility score (0.173) was achieved using Model43. Other investigated models performed similarly, high AUROC, AUPRC and accuracy, and low Utility score; or low AUROC, AUPRC and accuracy, and higher Utility score was observed in all investigated models, namely decision trees, SVM, naive Bayes, and ensemble-based models.

Highest F-measure and MCC scores using decision tree models were achieved for models that showed the highest Utility scores. F-measure score of Model14 was 0.133 (Utility—0.242), score of Model11 was 0.131 (Utility—0.232). MCC score of Model14 was 0.143; the score of Model11 was 0.14. Lowest F-measure and MCC scores were achieved using Model2: F-measure—0.011, MCC—0.018. However, the lowest Utility score (-0.296) using decision tree models was achieved using Model5. F-measure score of Model5 was 0.04, MCC—0.057.

Naive Gaussian Bayes models achieved lowest F-measure (0.062) and MCC (0.085) scores when Model22 (Utility score—0.097) was scored, highest F-measure (0.099) and MCC (0.122) scores using Model24 (Utility score—0.15).

Model32 trained using SVM achieved F-measure (0.108) and MCC (0.104) score, and scored 0.144 using Utility performance metric. However, Model31's Utility score (0.1515) was slightly higher, while F-measure (0.099) and MCC (0.1) score lower. Low F-measure (0.081) and MCC (0.087) scores also showed low Utility scores—0.129, for Model33.

Comparably low F-measure (0.025) and MCC (0.02) scores using ensemble learners were achieved using Model44 (Utility score—0.008). Model43 with a high Utility score (0.173) demonstrated a high F-measure (0.115) and MCC (0.113) scores.

Table 2. Results of the investigated models, baseline scores, and five highest Utility scores on hidden test set from the challenge. Our Investigated models labeled from 1 to 14 were based on decision trees. Models labeled from 21 to 25 were based on naive Bayes algorithm. SVM based models were labeled from 31 to 34. Models labeled from 41 to 44 were using ensemble methods with hyperparameter search. Best performing models, having the highest Utility score, were highlighted.

Model	AUROC	AUPRC	Accuracy	F-Measure	MCC	Utility
All positive	0.000	0.000	0.014	0.029	0.000	−1.059
All negative	0.500	0.014	0.986	0.000	0.000	0.000
Random performance	**0.125**	**0.007**	**0.500**	**0.027**	**0.000**	**−0.529**
Hong et al. [31]	**0.060**	**0.003**	**0.937**	**0.094**	N/A	**0.193**
Murugesan et al. [32]	0.256	0.006	0.962	0.113	N/A	0.182
Narayanaswamy et al. [33]	0.701	0.069	0.881	0.059	N/A	0.062
Alfaras et al. [34]	0.702	0.078	0.877	0.058	N/A	0.055
Deogire [35]	0.586	0.016	0.984	0.048	N/A	0.017
Proposed method	**0.307**	**0.009**	**0.934**	**0.130**	**0.143**	**0.245**
when trained on both datasets	0.291	0.009	0.934	0.140	0.158	0.276
Model1	0.492	0.014	0.984	0.024	0.028	0.010
Model2	**0.497**	**0.014**	**0.985**	0.011	0.018	0.004
Model3	0.491	0.014	0.984	0.028	0.033	0.012
Model4	0.379	0.011	0.945	0.095	0.090	0.124
Model5	0.058	0.003	0.486	0.040	0.057	−0.296
Model6	0.097	0.005	0.562	0.040	0.051	−0.242
Model7	0.357	0.011	0.950	0.126	0.125	0.184
Model8	0.314	0.010	0.930	0.118	0.129	0.220
Model9	0.278	0.009	0.901	0.100	0.119	0.216
Model10	0.321	0.010	0.914	0.091	0.098	0.157
Model11	0.321	0.010	0.940	0.131	0.140	0.232
Model12	0.309	0.009	0.931	0.124	0.136	0.233
Model13	0.309	0.009	0.931	0.124	0.136	0.233
Model14	0.313	0.009	0.937	**0.133**	**0.143**	**0.242**
Model21	0.226	0.008	0.828	0.070	0.092	0.133
Model22	**0.241**	**0.009**	**0.874**	0.062	0.085	0.097
Model23	0.23	0.008	0.831	0.069	0.090	0.129
Model24	0.201	0.006	0.785	**0.099**	**0.122**	**0.150**
Model25	0.202	0.006	0.790	0.092	0.118	0.143
Model31	0.353	0.011	0.935	0.099	0.100	**0.151**
Model32	0.375	0.011	0.949	**0.108**	**0.104**	0.144
Model33	0.320	0.010	0.904	0.081	0.087	0.129
Model34	**0.392**	**0.012**	**0.956**	**0.109**	0.101	0.130
Model41	0.413	0.012	0.955	0.084	0.072	0.082
Model42	0.358	0.011	0.937	0.100	0.099	0.124
Model43	0.347	0.010	0.939	**0.115**	**0.113**	**0.173**
Model44	**0.489**	**0.014**	**0.981**	0.025	0.020	0.008

4. Discussion

The highest Utility score was achieved using our proposed method, which divided patients based on their length of stay and then the appropriate model was applied. Additionally, decision trees with a low number of nodes achieved high Utility scores when ICU length of stay was included as a branch of decision tree. Therefore, we believe that future models should be developed based on ICU-stay time. For example, one model predicting recently hospitalized ICU patients, another would

be used if a patient's ICU length of stay reaches a certain length of time. Also, this approach can be implemented using three or more temporal divisions. This finding of our investigation is supported by Lauritsen [36], Vincent [37] and Shimabukuro [38] papers. Each intervention, vital measurement, intravenous therapy, and duration of stay in general increases a chance of infection—a direct cause of sepsis.

Regarding the dataset, other papers tackled this problem and proposed methods, which were trained on both datasets, and officially scored on hidden set C. Dataset C is not available anymore. Therefore, one must find other means to compare the results with the challenge score. Most of the challengers performed well on known hospital systems, obtaining Utility scores of about 0.4. However, Utility scores for the hidden hospital systems were low [5]. One author suggested evaluating the proposed methodology using one dataset and testing it on another [25]. Our achieved Utility score was for the known dataset, but the hidden percentage of data was 0.276 when trained on 75% of the records. This shows that our proposed model is robust to overtrain.

We assume that the Utility score can be improved a little by finding better value for classification cost, where a true positive prediction reward would be multiplied somewhere between 20 and 30. However, this would fit the data and would not solve the general problem of the Challenge. Therefore, we recommend using an arbitrary value between 20 and 30 to increases the robustness of the system.

MCC and F-measure scores gave similar results, which increases and decreases with the Utility score. However, the bounds of MCC are from -1 to 1, while the F-measure is from 0 to 1. The bounds of the Utility score are from -2 to 1. We support the idea of using the Utility score as a metric for this dataset. Moreover, we showed that the MCC and F-measure are effective metrics for this problem, while other traditional metrics AUROC, AURRC and Accuracy are misleading for a highly unbalanced dataset. Additionally, due to the nature of the Utility score, results can be difficult to interpret, as Roussel et al. pointed out in their work [39].

Investigated decision trees achieved Utility score of 0.242, AUROC score of 0.313, and MCC score of 0.143 on hidden set. Models with such results are far from applicable to the clinical setting. Additionally, our investigation showed that increasing AUROC and accuracy usually leads to decreased Utility, F-measure, and MCC scores. Moreover, accuracy is high for all investigated models. Accuracy can be miss-leading when interpreting models, results for this kind of highly unbalanced data, and a large number of negatives [22]. When developing methods for this kind of problem, one needs to be careful; the accuracy of 98.2% can be achieved just by guessing all rows as negative. We showed that balancing data reduces AUROC, accuracy scores and improves F-measure, MCC, and Utility scores.

There are many models to experiment with, for example, k-Nearest Neighbor (kNN) and Long Short-Term Memory (LSTM) models were not tested in our work. LSTM models are more difficult to configure to use them effectively. Additionally, LSTM tends to overfit the data. Moreover, even if one successfully tackles the overfitting problem, there is still another downside, which is more important in the current state of the early sepsis prediction problem. The developed model may be hard to interpret and would not reveal much insight into data [40]. The clustering of unbalanced data (including Sepsis-related records) may give promising results for sepsis prediction. However, kNN overfits data with large variances [41]. On the other hand, a trained kNN model having 1000 or 2000 clusters to represent the data can be expected to be robust. In general, we believe results using these models can be promising, and we encourage future works exploring LSTM and kNN model capabilities.

It is notable that investigated models do not differ significantly in Utility score if a number of features is reduced. This shows that some features are not useful for the model. On the other hand, our proposed features were relatively simple. We believe that more advanced features are needed to solve the early detection of the sepsis problem. Using advanced features should improve the score. However, feature engineering is a difficult, time-consuming process, which also requires

understanding the nature of the data. In this paper, we provide many insights into the nature of the data, different scoring metrics, advantages of various models, and feature combinations.

We believe that the results of our investigation presented in this paper will benefit the fundamental need of early sepsis prediction and will answer some basic questions about the limits of early detection. Our results should benefit the search for advanced combinations of features, ease the use of machine learning tools. With meaningful insights peer researchers can apply advanced feature engineering techniques and develop more sophisticated and robust models in order to reach reliable results. Reaching better results is available through the use of combined models and handcrafted features [42], thus, further contributing to this field. The main challenges of this problem, as we revealed, are—the highly unbalanced dataset, the high number of missing data, simple features calculated using vitals does not have enough predictive power, proposed solutions are prone to overtrain. Adjusting the classification cost function helps to address the latter problem. In addition, the insights and conclusions of our experiment may benefit not only machine learning specialists, researchers, but also ICU personnel and scientists in the medical field.

5. Conclusions

In this study, we provide a comparison of several alternative methods for early sepsis prediction. The performance of the investigated models was based on the Utility score metric. Our selected models and insights show how to deal with unbalanced data and with a large number of missing values.

The results, obtained during an experimental investigation, are based on publicly available data containing 40,336 records with 1,407,716 of rows and 40 dimensions. Results showed:

1. Our proposed method, using two separate ensemble models, based on length of stay in the ICU, performed better than other tested models when using vital and demographic data to calculate simple features.
2. Adjusting classification cost function improves the Utility score of the tested models. Best results, on the investigated dataset, were achieved when the reward of true positive prediction was increased 20 times.
3. Feature ranking, using PCA, applied for our proposed features does not always improve Utility score. Utility score changes, when reducing the number of features based on the investigated model. In some models, such as Naive Gaussian, reducing the number of features improved the Utility score.
4. Performance metrics AUROC, AUPRC, and accuracy are not suitable for this highly unbalanced dataset. Additionally, these metrics do not reflect the Utility score. These metrics can be high for models with low Utility scores. Dealing with the early sepsis prediction problem, one should not apply these performance metrics. On the contrary, F-measure and MCC performance metrics reflect the Utility score.
5. High Utility score was obtained using decision tree models limited to 50 and fewer splits. All investigated decision trees chose ICULOS—ICU length of stay, as an important feature. Additionally, reducing the number of tree splits up to 4, and 1 further increased the Utility score. Utility score of 0.242 was achieved using only ICULOS as a single feature for the decision tree model.

Author Contributions: Conceptualization and methodology, all authors; validation, V.A. and D.T.; Analysis, A.S. and V.A.; writing—original draft preparation, A.S. and V.A.; project administration, A.S.; supervision, D.P.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Singer, M.; Deutschman, C.S.; Seymour, C.W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G.R.; Chiche, J.D.; Coopersmith, C.M.; et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* **2008**, *10*, 142–149. [CrossRef]
2. Vincent, J.L.; Marshall, J.C.; Ñamendys-Silva, S.A.; François, B.; Martin-Loeches, I.; Lipman, J.; Reinhart, K.; Antonelli, M.; Pickkers, P.; Njimi, H.; et al. Assessment of the worldwide burden of critical illness: The intensive care over nations (icon) audit. *Lancet Respir. Med.* **2014**, *2*, 380–386. [CrossRef]
3. Fleischmann, C.; Scherag, A.; Adhikari, N.K.; Hartog, C.S.; Tsaganos, T.; Schlattmann, P.; Angus, D.C.; Reinhart, K. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *Am. J. Respir. Crit. Care Med.* **2016**, *193*, 259–272. [CrossRef]
4. Reinhart, K.; Daniels, R.; Kissoon, N.; Machado, F.R.; Schachter, R.D.; Finfer, S. Recognizing sepsis as a global health priority—A who resolution. *N. Engl. J. Med.* **2017**, *377*, 414–417. [CrossRef]
5. Reyna, M.A.; Josef, C.S.; Jeter, R.; Shashikumar, S.P.; Westover, M.B.; Nemati, S.; Clifford, G.D.; Sharma, A. Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. *Crit. Care Med.* **2020**, *48*, 210–217. [CrossRef]
6. 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] [PubMed]
7. Vicar, T.; Hejc, J.; Novotna, P.; Ronzhina, M.; Smisek, R. Sepsis Detection in Sparse Clinical Data Using Long Short-Term Memory Network with Dice Loss. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
8. Sweely, B.; Park, A.; Winter, L.; Liu, L.; Zhao, X. Time-Padded Random Forest Ensemble to Capture Changes in Physiology Leading to Sepsis Development. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
9. He, Z.; Chen, X.; Fang, Z.; Yi, W.; Wang, C.; Jiang, L.; Pan, Y. Early Sepsis Prediction Using Ensemble Learning with Features Extracted from LSTM Recurrent Neural Network. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
10. Li, X.; Kang, Y.; Jia, X.; Wang, J.; Xie, G. TASP: A Time-Phased Model for Sepsis Prediction. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
11. Nejedly, P.; Plesinger, F.; Viscor, I.; Halamek, J.; Jurak, P. Prediction of Sepsis Using LSTM with Hyperparameter Optimization with a Genetic Algorithm. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
12. Yang, M.; Wang, X.; Gao, H.; Li, Y.; Liu, X.; Li, J.; Liu, C. Early Prediction of Sepsis Using Multi-Feature Fusion Based XGBoost Learning and Bayesian Optimization. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
13. Morrill, J.; Kormilitzin, A.; Nevado-Holgado, A.; Swaminathan, S.; Howison, S.; Lyons, T. The Signature-Based Model for Early Detection of Sepsis from Electronic Health Records in the Intensive Care Unit. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
14. Anda Du, J.; Sadr, N.; de Chazal, P. Automated Prediction of Sepsis Onset Using Gradient Boosted Decision Trees. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
15. Hammoud, I.; Ramakrishnan, I.; Henry, M. Early Prediction of Sepsis Using Gradient Boosting Decision Trees with Optimal Sample Weighting. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
16. Singh, J.; Oshiro, K.; Krishnan, R.; Sato, M.; Ohkuma, T.; Kato, N. Utilizing Informative Missingness for Early Prediction of Sepsis. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
17. Tran, L.; Shahabi, C.; Nguyen, M. Representation Learning for Early Sepsis Prediction. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.

18. DeCarlo, L.T. On the meaning and use of kurtosis. *Psychol. Methods* **1997**, *2*, 292. [CrossRef]
19. Dehmer, M. Information processing in complex networks: Graph entropy and information functionals. *Appl. Math. Comput.* **2008**, *201*, 82–94. [CrossRef]
20. Hamaker, E.L.; Ryan, O. A squared standard error is not a measure of individual differences. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6544–6545. [CrossRef]
21. Tanha, J.; van Someren, M.; Afsarmanesh, H. Semi-supervised self-training for decision tree classifiers. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 355–370. [CrossRef]
22. Hu, B.; Wang, J.; Zhu, Y.; Yang, T. Dynamic Deep Forest: An Ensemble Classification Method for Network Intrusion Detection. *Electronics* **2019**, *8*, 968. [CrossRef]
23. Chen, Y.; Lu, L.; Yu, X.; Li, X. Adaptive Method for Packet Loss Types in IoT: An Naive Bayes Distinguisher. *Electronics* **2019**, *8*, 134. [CrossRef]
24. Gu, B.; Quan, X.; Gu, Y.; Sheng, V.S.; Zheng, G.S. Chunk incremental learning for cost-sensitive hinge loss support vector machine. *Pattern Recognit.* **2018**, *83*, 196–208. [CrossRef]
25. Biglarbeigi, P.; McLaughlin, D.; Rjoob, K.; Abdullah, A.; McCallan, N.; Jasinska-Piadlo, A.; Bond, R.; Finlay, D.; Ng, K.Y.; Kennedy, A.; et al. Early Prediction of Sepsis Considering Early Warning Scoring Systems. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
26. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
27. Boyd, K.; Eng, K.H.; Page, C.D. Area under the precision-recall curve: Point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2013.
28. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
29. Rousseau, R. The F-measure for research priority. *J. Data Inf. Sci.* **2018**, *3*, 1–18. [CrossRef]
30. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
31. Hong, S.; Shang, J.; Wu, M.; Zhou, Y.; Sun, Y.; Chou, Y.H.; Song, M.; Li, H. Early Sepsis Prediction with Deep Recurrent Reinforcement Learning. *Physionet Chall.* **2019**. Available online: https://docs.google.com/spreadsheets/d/1PPQY0SdguwCx_CxbR1BYlkh0dwpINlhEFxejc10xwgM/edit?fbclid=IwAR3psdL1QQ_PxlukPT89fE-v0ZVFgLdax11mrAeQwkdCO9WXeEKFfn8ek2o#gid=0 (accessed on 5 May 2020).
32. Murugesan, I.; Murugesan, K.; Balasubramanian, L.; Arumugam, M. Interpretation of Artificial Intelligence Algorithms in the Prediction of Sepsis. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
33. Narayanaswamy, L.; Garg, D.; Narra, B.; Narayanswamy, R. Machine Learning Algorithmic and System Level Considerations for Early Prediction of Sepsis. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
34. Alfaras, M.; Varandas, R.; Gamboa, H. Ring-Topology Echo State Networks for ICU Sepsis Classification. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
35. Deogire, A. A Low Dimensional Algorithm for Detection of Sepsis From Electronic Medical Record Data. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
36. Lauritsen, S.M.; Kalør, M.E.; Kongsgaard, E.L.; Lauritsen, K.M.; Jørgensen, M.J.; Lange, J.; Thiesson, B. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Early Detect. Sepsis Util. Deep Learn. Electron. Health Rec. Event Seq.* **2020**, *104*, 101820. [CrossRef]
37. Vincent, J.L. The clinical challenge of sepsis identification and monitoring. *PLoS Med.* **2016**, *13*, e1002022. [CrossRef]
38. Shimabukuro, D.W.; Barton, C.W.; Feldman, M.D.; Mataraso, S.J.; Das, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. *BMJ Open Respir. Res.* **2017**, *4*, e000234. [CrossRef] [PubMed]
39. Roussel, B.; Behar, J.; Oster, J. A Recurrent Neural Network for the Prediction of Vital Sign Evolution and Sepsis in ICU. In Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, 8–11 September 2019.
40. Wan, R.; Mei, S.; Wang, J.; Liu, M.; Yang, F. Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. *Electronics* **2019**, *8*, 876. [CrossRef]

41. Mullick, S.S.; Datta, S.; Das, S. Daptive Learning-Based *k*-Nearest Neighbor Classifiers With Resilience to Class Imbalance. *IEEE Trans. Neural Networks Learn. Syst.* **2018**, *29*, 5713–5725.
42. Sawada, Y.; Sato, Y.; Nakada, T.; Yamaguchi, S.; Ujimoto, K.; Hayashi, N. Improvement in Classification Performance Based on Target Vector Modification for All-Transfer Deep Learning. *Appl. Sci.* **2019**, *9*, 128. [CrossRef]