

# Mortality prediction for ICU patients with individualized single classification method

Xin Ma\* Xiaosu Guo\* Youqing Wang\*\*\*

\* College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China (e-mail: 838154269@qq.com).

\*\* College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China (\*corresponding author e-mail: wang.youqing@ieee.org).

---

**Abstract:** In intensive care units (ICUs), mortality prediction is crucial to help doctors select appropriate diagnosis and treatment methods and reasonably allocate medical resources. However, the scoring system currently used by hospitals may not be suitable for every patient. Therefore, personalized diagnosis is trending because it responds to a wide range of needs. This study proposes a new combination of dynamic time warping (DTW) and one-class extreme learning machine (ELM) to improve the mortality prediction accuracy, termed as one-class DTW-ELM, where DTW is used to find similar cases for new patients, and while one-class ELM is adopted for fast and accurate model building. The testing results on the real physiological data from PhysioNet demonstrate that the area under curve (AUC) index of one-class DTW-ELM model is 0.9739, with the Lift of 8.1081 and the G-mean of 0.8137. The accuracy rate is 0.9583, with specificity and sensitivity of 1 and 0.6622, respectively. As mentioned, one-class DTW-ELM can accurately predict the future physiological state of a patient by using early physiological parameters.

*Keywords:* Mortality prediction, ICU patients, personalized diagnosis, dynamic time warping (DTW), extreme learning machine (ELM)

---

## 1. INTRODUCTION

As medical and nursing technology continuously improves, intensive care units (ICUs) progressively play a pivotal role in modern hospitals. Nowadays, numerous models have been established to assess patient status based on physiological indicators (Valenza et al., 2018). Well-known methods based on generalized linear models are commonly used for mortality prediction problems, such as acute physiology and chronic health evaluation, simplified acute physiology score, and mortality probability model. Given the in-depth study on technology by many scholars and the increasing amount of clinical data in ICUs, many researchers have adopted data-driven learning methods to implement mortality prediction (Ma et al., 2020). In particular, machine learning involves various non-linear modeling methods, and only relies on physiological measurements to improve the accuracy of mortality prediction, such as neural networks (Ding et al., 2018), support vector machine (SVM) (Azar et al., 2014), and naive Bayesian model.

The mortality prediction of ICU patients not only assists doctors in making clinical decisions during the diagnosis stage but also optimizes the allocation of medical resources. In fact, because each ICU patient has a specific

status, all the data may not provide an exceedingly targeted diagnosis and treatment program for individual patients. However, due to the advent of intelligent medicine, personalized treatment for patients is achievable.

In this study, specificity of ICU patients is fully considered, and the patients personalized diagnosis is achieved based on the idea that similar inputs yield similar outputs. Thus, DTW is selected as a personalized search tool to ensure that the local model is accurately targeted. In particular, one-class ELM is used as a classifier. What should be emphasized is that in the case of severe data imbalance, deaths accounted for a low percentage of total cases are regarded as outliers and the one-class classification method is commonly in accordance with the actual situation and has achieved ideal results.

Furthermore, the proposed algorithm is verified by means of three sets of comparative experiments. The experimental datas come from PhysioNet, which contained real physiological information of 4,000 patients, 554 of whom survived and 3446 died in hospital. The experimental results are 0.9739 for the area under curve (AUC) index of the one-class DTW-ELM model, 8.1081 for Lift, 0.8137 for G-mean, 0.9583 for accuracy rate, 1 for specificity, and 0.6622 for sensitivity.

Overall, the main findings and contributions of this study are as follows: (1) A personalized model based on patients time-series data is proposed. Through this model, the state of inpatients can be judged accurately. (2) Personalized

---

\* This study was supported by the National Natural Science Foundation of China under Grant 61822308, Shandong Province National Science Foundation under Grant JQ201812, and the Research Fund for the Taishan Scholar Project of Shandong Province of China.

search based on time series can help doctors find similar cases. (3) Turning the binary classification problem into a single classification problem, using only the physiological indicators for the first 48 hours after admission to the ICU, can more accurately predict the patient's status.

The rest of this paper is organized as follows. Section 2 summarizes data collection, data interpolation, and principal component analysis. Section 3 introduces one-class DTW-ELM. Section 4 presents illustrative experiments and results. Finally, Section 5 draws conclusions.

## 2. DATA PROCESSING

### 2.1 Data Collection

The physiological data of ICU patients are collected from the PhysioNet website (Johnson et al., 2016). The website offers free access to complex physiological signals to help scholars conduct in-depth research. The scientific and rigorous information of this authoritative website has been widely validated.

Herein, total 4,000 real ICU patient records are collected, which includes 554 dead patients and 3,446 surviving patients. The physiological parameters of each ICU patient first 48 hours after admission to the ICU were recorded truthfully. The data collected generally consist of three parts, which includes basic information about hospitalization, 36 physiological variables, and final state of the patient (0 = survivor, 1 = deceased), as shown in Table 1.

Table 1. PHYSIOLOGICAL DATASET.

RecordID=132540		Final State: 0			
Time	HR	Temp	GCS	...	NIDiasABP
0:00	-1	-1	-1	...	-1
0:42	-1	-1	-1	...	-1
1:11	88	35.2	-1	...	67
1:26	88	35.1	3	...	69
...	...	...	...	...	...
46:15	-1	-1	-1	...	-1
47:11	65	37.1	15	...	49

'-1' denotes the missing value.

However, the sampling frequency of these physiological parameters is different, not all of them are suitable for accurately predicting the future state of patients. Among these parameters, some are selected for modeling. Therefore, 15 physiological parameters with the most complete information are selected, as shown in Table 2.

### 2.2 Data interpolation

Missing data represent a problem in physiological parameters. To interpolate the missing data while simultaneously retaining the time information to the greatest extent, irregular ICU physiological variables for each sampling period are converted to a time series of periodic sampling, sampled every two hours (McMillan et al., 2012).

Data from the first two days of ICU admission are used. For each indicator of the patient, it is divided into 24 segments by time and one value per segment. If there are more than one detection value in a certain section of the

Table 2. LIST OF 15 SELECTED PHYSIOLOGICAL PARAMETERS AND THEIR ABBREVIATIONS.

Abbreviation	Name
HR	Heart rate (bpm)
RespRate	Respiration rate (bpm)
GCS	Glasgow coma index<3-15>
DiasABP	Invasive diastolic arterial blood pressure (mmHg)
NIMAP	Non-invasive mean arterial blood pressure (mmHg)
FiO2	Fractional inspired O2 <0-1>
NIDiasABP	Non-invasive diastolic arterial blood pressure (mmHg)
SysABP	Invasive systolic arterial blood pressure (mmHg)
Temp	Temperature (°C)
PaCO2	Partial pressure carbon dioxide
MAP	Invasive mean arterial blood pressure (mmHg)
MechVent	Mechanical ventilation respiration (0: false, or 1: true)
PH	Arterial pH <0-14>
NISysABP	Non-invasive systolic arterial blood pressure (mmHg)
Urine	Urine output (mL)

index, the average of these detection values is taken. If there is no detection value in one segment of the index, the average value of the patient in the other segments of the index is used to interpolate. Some patients do not have the test data of one index, then interpolate the data in these 24 segments with the overall average of other patients. The total number of extracted medical information features is 360.

Herein, the dimension of extracted features is high. Before features input into the model, principal component analysis (PCA) algorithm can reduce the amount of calculation (Lou et al. 2017), extract the main information and avoid the influence of useless information (Wang et al., 2018). For the remaining information, 267 principal components are selected, including 99 percent of cumulative variance contribution rate, which would be the input data of one-class classification DTW-ELM model.

## 3. ONE-CLASS DTW-ELM

ICU patients are highly specific, so using all the data to model patients does not meet the individual needs of patients, as shown in Fig. 1. Based on the idea of getting similar output based on similar inputs, the ICU patient personalized search becomes the key to solving this problem. Herein, DTW (Gogolou et al., 2019) is selected as a personalized search tool to ensure that the local model is more targeted.

Simultaneously, the classification accuracy of the model is also a major factor. The common classifier adopts a binary classification algorithm to learn the laws of two types of data through the training set. However, the physiological parameters of ICU patients are extremely special. A serious imbalance exists in the data, and the ratio of survival to death is about 7:1. Those who can survive in hospital conform to certain rules, unlike those who die.

Interquartile range (IQR) is a quantity used to represent the degree of data dispersion in robust statistics. We know that the larger the value of IQR, the more scattered

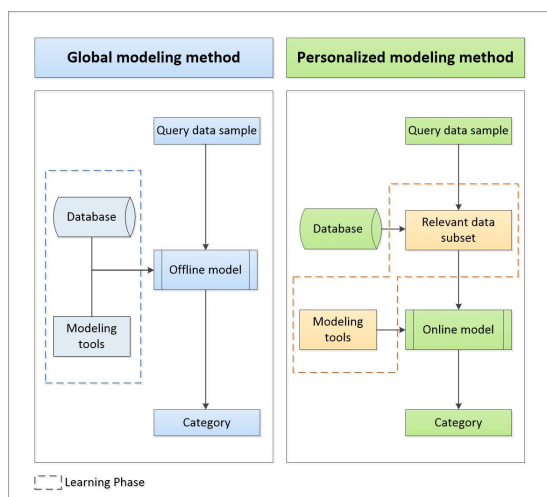


Fig. 1. Mechanism comparison of conventional model and personalized model.

the data, and the smaller the value of IQR, the more centralized the data. The X-axis corresponds to each physiological parameter, and the y-axis represents the difference of the IQR between the dead and the surviving patients in this physiological parameter. The blue line indicates that the difference is 0 and the IQR is equal. If the red dot is higher than the blue line, it means that the IQR of the dead patient is larger than that of the surviving patient. As shown in the Fig 2, points above the blue line are much more than points below the line. This indicates that most IQR are less in survivors than in dead patients. This shows that the distribution of physiological parameters of surviving patients is more centralized, more in line with certain rules, while the distribution of physiological parameters of dead patients is more scattered and more disordered. So we choose more centralized survival data as training set and use one class classification method to model. If the classifier learns two kinds of data, the result will become inaccurate.

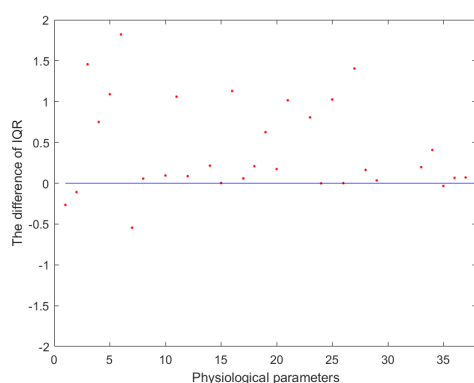


Fig. 2. The difference of the IQR between the physiological parameters of the two kinds of data.

Accordingly, one-class ELM is picked as the local modeling tool instead of ELM (Huang et al., 2012), and transforms classification problems into outlier detection problems. This method provides an effective way to solve the problem of low prediction accuracy due to the unbalanced distribution of training samples. In particular, only data of

patients who survive in hospital are used as training set, and death data are considered abnormal when classified.

One-class DTW-ELM modeling algorithm is described as follows:

Step 1. Total 600 records are randomly selected as the testing samples, and 2,900 survival records are selected from the remaining data as a training set.

Step 2. Each time a test sample is read, then the 200 most similar samples are selected as a personalized training set for use as modeling data in the next step.

Step 3. For the current testing patient, a one-class ELM model is established using 200 most similar samples.

Step 4. The current test data is input into the established model to obtain the predicted result.

Step 2 is repeated if new testing sample data are available.

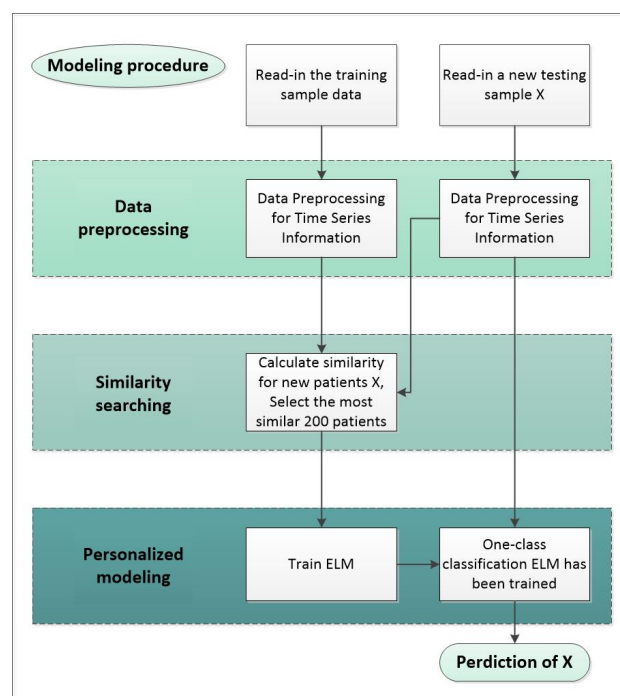


Fig. 3. One-class DTW-ELM model procedure.

Fig. 3 shows the working process of the one-class DTW-ELM model. In practice, the test data is read in order, and the one-class ELM model is built for each patient by its most similar samples. New models are built in real time, and old models are abolished. The idea for this process is similar to that of diagnosing diseases. When a new patient is encountered, the doctor searches for similar cases in the brain according to the patient's symptoms and provides the final diagnosis and treatment plan.

Compared with the traditional modeling method, the one-class DTW-ELM model is simpler and more accurate. Simultaneously, it is more in line with the real diagnosis process of doctors.

#### 4. EXPERIMENTS AND RESULTS

Herein, we use three sets of experiments to verify the superiority of one-class DTW-ELM in ICU patients physiological parameters. And used several indicators to prove the validity of the classifier. Traditional indicators are widely used, however, the overall accuracy does not apply to the datasets with imbalance. Because when a few classes are misclassified, the overall accuracy rate may still have a good indicator. In this study, imbalanced problem also evidently exists. Surviving patients account for the majority, and death patients account for a minority. In the medical field, “positive” and “negative” are often used to represent two categories; the former represents death and the latter represents survival.

Confusion matrix is often used to judge the performance of classification algorithm. TP, FP, FN, and TN indicate the number of samples of true positive, false positive, false negative, and true negative, respectively. Moreover, many other indices are available, such as

$$\text{sensitivity} : Sn = \frac{TP}{TP + FN} \quad (1)$$

$$\text{specificity} : Sp = \frac{TN}{TN + FP} \quad (2)$$

$$G\text{-mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (3)$$

$$\text{Lift} = \frac{TP}{TP + FP} \bigg/ \frac{TP + FN}{\text{all}} (\text{Lift}_{\min} = 1) \quad (4)$$

$$\text{Acc} = (TP + TN) / \text{all} \quad (5)$$

where sensitivity refers to the proportions of true positive samples that can be correctly detected, and specificity indicates that the proportion of true negative samples are not wrongly misclassified; G-mean refers to the geometric mean of those accuracies, which indicates the final result to measure the functionality of a classifier; and Lift describes the performance improvement compared with the non-use model, the minimum value of which is 1. Receiver operating characteristics (ROC) graph and AUC are also used for comparison in the experiment. The area under a ROC curve should be at least 0.75 for good prediction. To ensure the fairness of the experiment, the test set of each model is the same 600 records. The survival and death records used in the training set are also the same.

##### 4.1 Experiment 1: Comparison with unsupervised algorithm

In this unsupervised experiment, the classical K-means clustering algorithm and the novel isolation forest algorithm (iForest) are selected for comparison. For fair comparison, unsupervised algorithms use only 600 test set data.

The core of K-means is to use k initial centroids as the clustering category and repeat iteration until the algorithm converges. In this paper, according to the experimental needs, k=2.

The iForest algorithm does not describe the normal point but is dedicated to isolating the abnormal point. The iForest has a efficient strategy for judging outliers. The anomaly point here is defined as an outlier that is easily isolated. The algorithm does not require training to

recursively segment the dataset until all sample points are isolated. Under this stochastic segmentation strategy, normal points are not easily segmented, while outliers usually have shorter paths. In this set of experiments, the threshold was set to 0.5.

In specificity, Lift, G-mean, accuracy and AUC, one-class DTW-ELM achieves the best results. In sensitivity, k-means works best, but its specificity and accuracy are too low, and the classification effect is not great. The iForest do not perform very well in all indicators, as shown in Table 3. ROC is presented in Fig. 4.

The results show that one-class DTW-ELM works better than the usual unsupervised methods.

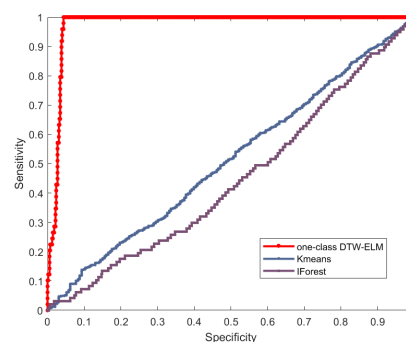


Fig. 4. ROC curves for one-class DTW-ELM, K-means and iForest.

##### 4.2 Experiment 2: Comparison with traditional binary classification algorithm

In this group of experiments, except one-class DTW-ELM, all the other models are binary classification methods. In the training of the model, besides 2,900 survival records, 400 death records were added proportionally to build the training set. Both DTW-ELM and DTW-SVM are binary classification algorithms that use personalized search. Similarly to one-class DTW-ELM, the closest 200 cases are selected for modeling.

At the same time, we added a set of comparative experiments to eliminate the impact of data imbalance on the data of binary training set. The SMOTE algorithm is used to increase the death data by 2500, so that the number of two kinds of samples in the training set is 2900. And the test set is the same for all algorithms.

In the contrast experiment of unbalanced training set, the results show that the sensitivity of SVM and DTW-ELM is equal to 0, and a serious problem occurs with failing reports, causing the G-mean to be 0, as shown in Table 4. For SVM, TP FP is zero, and the classifier has no discriminant ability. It regards all patients as viable, which leads to the absence of its AUC value. In such cases, the ROC curves are no longer drawn. ROC is presented in Fig. 5.

In another set of comparisons, we use SMOTE algorithm to balance training set (Chawla et al., 2011). SMOTE algorithm does not simply copy minority samples, but seeks to balance the class distribution of training samples

Table 3. COMPARISON WITH UNSUPERVISED ALGORITHM.

	TP	TN	FP	FN	Sn	Sp	Lift	Gmean	Acc	Auc
one-class DTW-ELM	49	526	0	25	0.6622	1	8.1081	0.8137	0.9583	0.9757
K-means	56	175	351	18	0.7568	0.3327	1.1156	0.5018	0.3850	0.5117
iForest	9	438	88	65	0.1216	0.8289	0.7523	0.3175	0.7417	0.4452

Table 4. COMPARISON WITH BINARY CLASSIFICATION ALGORITHMS.

	TP	TN	FP	FN	Sn	Sp	Lift	Gmean	Acc	Auc
one-class DTW-ELM	49	526	0	25	0.6622	1	8.1081	0.8137	0.9583	0.9757
SVM	0	526	0	74	0	1	NaN	0	0.8767	NaN
ELM	1	523	3	73	0.0135	0.4771	2.0270	0.1159	0.8733	0.4771
DTW-SVM	0	513	13	74	0	0.97529	0	0	0.8550	0.5841
DTW-ELM	49	526	0	25	0.5270	0.5741	1.2023	0.5501	0.5683	0.5068

”NaN” represents the absence of this indicator due to all missed reports.

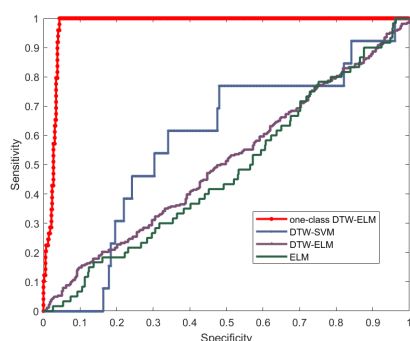


Fig. 5. ROC curves for one-class DTW-ELM, DTW-SVM and DTW-ELM and ELM.

by generating a large number of new samples with certain strategies.

The result shows that all indicators of one-class DTW-ELM are in the best position in this comparison, as shown in Table 5. ROC is presented in Fig. 6. SMOTE-SVM and SMOTE-DTW-SVM have the same perfect specificity, but their sensitivity is 0. In this case, no matter how high the accuracy is, the classification effect is not ideal.

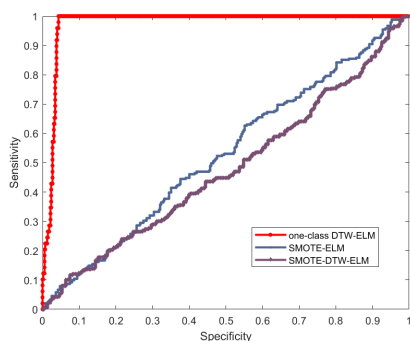


Fig. 6. ROC curves for one-class DTW-ELM, SMOTE-ELM and SMOTE-DTW-ELM.

Therefore, compared with the binary classification, one-class DTW-ELM performs well. This result also indicates that the data may not be suitable for classification by common binary classification methods.

### 4.3 Experiment 3: Comparison with one-class classification algorithm

This set of Comparisons randomly select the same 2,900 survival records as the train set and 600 records as the test set. Of these, 526 were survival records and 74 were death records.

The experiment results are obvious, and all the death records in one-class ELM model are not reported, as shown in Table 6. As both TP and FP are 0, the sensitivity of one-class ELM is 0. This causes the denominator of Lift’s formula to become 0, so Lift does not exist and is represented by NaN in the table. AUC functions in a similar manner. In this group of experiments, one-class DTW-ELM has the highest sensitivity, but its specificity is zero and its classification is invalid. The specificity and sensitivity of one-class DTW-ELM are 1 and 0.6622, respectively, which means that all survival records can be judged correctly.

One-class SVM is a classical single classification model, which is mostly used in outlier detection. Owing to the imbalance of records, a good idea is to regard death records as outliers. The comparison shows that one-class SVM can be used in classification, but the classification effect is not ideal, as shown in Table 6. ROC is shown in Fig. 7. In both specificity and sensitivity, one-class SVM is inferior to one-class DTW-ELM. In particular, for the AUC of one-class SVM, the index is even lower than 0.5. In terms

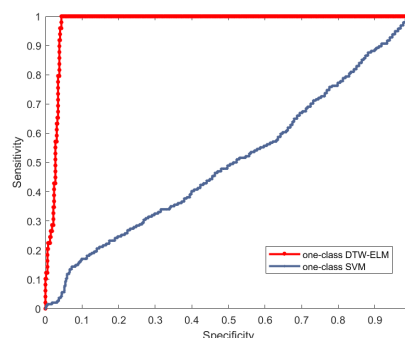


Fig. 7. ROC curves for one-class DTW-ELM and one-class SVM.

of various indicators, one-class DTW-ELM has obvious advantages. The experimental results show that one-class DTW-ELM still has obvious advantages over common single classification algorithms.

Table 5. COMPARISON WITH BINARY CLASSIFICATION ALGORITHMS (SMOTE).

	TP	TN	FP	FN	Sn	Sp	Lift	Gmean	Acc	Auc
one-class DTW-ELM	49	526	0	25	0.6622	1	8.1081	0.8137	0.9583	0.9757
SMOTE-SVM	0	526	0	74	0	1	NaN	0	0.8767	NaN
SMOTE-ELM	33	318	208	41	0.4459	0.6046	1.1102	0.5192	0.5850	0.5272
SMOTE-DTW-SVM	0	526	0	74	0	1	NaN	0	0.8767	NaN
SMOTE-DTW-ELM	42	251	275	32	0.5676	0.4772	1.0743	0.5204	0.4883	0.4772

"NaN" represents the absence of this indicator due to all missed reports.

Table 6. COMPARISON WITH ONE-CLASS CLASSIFICATION ALGORITHMS.

	TP	TN	FP	FN	Sn	Sp	Lift	Gmean	Acc	Auc
one-class DTW-ELM	49	526	0	25	0.6622	1	8.1081	0.8137	0.9583	0.9757
one-class ELM	0	526	0	74	0	1	NaN	0	0.8767	NaN
one-class SVM	33	365	161	41	0.4459	0.6939	1.3792	0.5563	0.6633	0.4984
one-class DTW-SVM	74	0	526	0	1	0	1	0	0.1233	NaN

"NaN" represents the absence of this indicator due to all missed reports.

## 5. CONCLUSION

By associating DTW for matching new patients with the same precedents having one-class ELM to set up a model fast and accurately, the proposed method accurately predicts mortality. The real data used in the study were obtained from the PhysioNet database, including 554 death and 3,446 survival records are used for ICU patients' mortality prediction. In terms of specificity, sensitivity, AUC, Lift, G-mean, and accuracy rate, one-class DTW-ELM achieves great performance compared with other algorithms. For the optimal tradeoff between specificity and sensitivity, the AUC index of one-class DTW-ELM model is 0.9757. The Lift is 8.1081 and the G-mean is 0.8137. The accuracy rate is 0.9583, with a specificity and sensitivity of 1 and 0.6622, respectively.

In summary, one-class DTW-ELM can use early physiological parameters to predict the future of patients. In addition, this method exhibits the outstanding performance of the early warning system and the reasonable allocation of medical resources, as well as consistency with the wide application of personalized medical treatment. In the future, other advanced methods could be combined with the proposed method, such as two-dimensional filtering (Zhao et al., 2017) and two-step PCA (Lou et al., 2018), to further improve the medical prediction performance and provide more effective treatment for patients.

## REFERENCES

- Valenza, G., Wendt, H., Kiyono, K., Hayano, J. and Watanabe, E. (2018). Mortality Prediction in Severe Congestive Heart Failure Patients with Multifractal Point-Process Modeling of Heartbeat Dynamics. *IEEE Transactions on Biomedical Engineering*, 65, 2345-2354.
- Ma, X., Si, Y., Wang, Z., Wang, Y. (2020). Length of stay prediction for ICU patients using individualized single classification algorithm. *Computer methods and programs in biomedicine*, 186, 105224.
- Ding, Y., Ma, X., Wang, Y. (2018). Health status monitoring for ICU patients based on locally weighted principal component analysis. *Computer methods and programs in biomedicine*, 281, 12-19.
- Azar, A. T., El-Said, S. A. (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, 24, 1163-1177.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.
- McMillan, S., Chia, C. C., Esbroeck, A. V., Rubinfeld, I., Syed, Z. (2012). ICU mortality prediction using time series motifs. *Computing in Cardiology*, 265-268.
- Lou, Z., Wang, Y. (2017) Multimode continuous processes monitoring based on hidden semi-Markov model and principal component analysis. *Industrial & Engineering Chemistry Research*, 56, 13800-13811.
- Wang, Y., Si, Y., Huang, B., Lou, Z. (2018). Survey on the Theoretical Research and Engineering Applications of Multivariate Statistics Process Monitoring Algorithms: 2008-2017. *The Canadian Journal of Chemical Engineering*, 96, 2073-2085.
- Gogolou, A., Tasandilas, T., Palpanas, T., Bezerianos, A. (2019). Comparing similarity perception in time series visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25, 523-533.
- Huang, G., Zhou, H., Ding, X., Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42, 513-529.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2011). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
- Zhao, D., Shen, D., Wang, Y. (2017). Fault diagnosis and compensation for two-dimensional discrete time systems with sensor faults and time-varying delays. *International journal of robust and nonlinear control*, 27, 3296-3320.
- Lou, Z., Shen, D., Wang, Y. (2018). Two-step principal component analysis for dynamic processes monitoring. *The Canadian Journal of Chemical Engineering*, 96, 160-170.