BenchCouncil Transactions on Benchmarks, Standards and Evaluations 1 (2021) 100004

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: https://www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Benchmarking feature selection methods with different prediction models on large-scale healthcare event data

Fan Zhang [a], Chunjie Luo [a,b,c], Chuanxin Lan [a], Jianfeng Zhan [a,b,c,*]

[a] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[c] International Open Benchmark Council (BenchCouncil)

ARTICLE INFO

Keywords:
Feature selection
Genetic algorithm
Deep neural networks
Healthcare prediction

ABSTRACT

With the development of the Electronic Health Record (EHR) technique, vast volumes of digital clinical data are generated. Based on the data, many methods are developed to improve the performance of clinical predictions. Among those methods, Deep Neural Networks (DNN) have been proven outstanding with respect to accuracy by employing many patient instances and events (features). However, each patient-specific event requires time and money. Collecting too many features before making a decision is insufferable, especially for time-critical tasks such as mortality prediction. So it is essential to predict with high accuracy using as minimal clinical events as possible, which makes feature selection a critical question. This paper presents detailed benchmarking results of various feature selection methods, applying different classification and regression algorithms for clinical prediction tasks, including mortality prediction, length of stay prediction, and ICD-9 code group prediction. We use the publicly available dataset, Medical Information Mart for Intensive Care III (MIMIC-III), in our experiments. Our results show that Genetic Algorithm (GA) based methods perform well with only a few features and outperform others. Besides, for the mortality prediction task, the feature subset selected by GA for one classifier can also be used to others while achieving good performance.

1. Introduction

Over the past decades, the Electronic Health Record (EHR) technique is developed; vast volumes of digital clinical data are generated, making it possible for Clinical Decision Support Systems (CDSSs) to make better decisions. For example, public databases such as MIMIC-III [1] have promoted the research in clinical predictions. Based on those databases, different severity scoring systems, traditional machine learning algorithms, and DNNs are developed and continuously improved to achieve better clinical prediction tasks such as patient mortality, disease classification, and length of hospital stay.

Traditional severity scores like Simplified Acute Physiology Score (SAPS-II) [2], the Sepsis-related Organ Failure Assessment (SOFA) [3], and Acute Physiology and Chronic Health Evaluation (APACHE) [4] are standard for mortality prediction in practice. Clinicians usually choose the patient-specific events they used based on their experience. Then a standard process is implemented. First, a severity score is calculated based on the relative events, usually measured within the first 24 h after ICU admission. Second, a simple model such as logistic regression is applied to the score to predict the final death probability.

Recent work shows that DNN and Super Learner (SL) algorithms perform better than single traditional classifiers and severity scoring systems [5–8]. To improve the predictive performance, many DNN models are developed. Purushotham et al. [6] proposed a Multimodal Deep Learning Model (MMDL) to process an extensive feature set, which consists of 141 features, and got very good predicting results. Harutyunyan et al. [7] proposed a multitask LSTM-based method to predict four clinical prediction tasks. In addition to DNN, SL is also studied extensively and shows promising results. Pirracchio et al. [5] provided and assessed the performance of the Super ICU Learner Algorithm (SICULA). Lee et al. [8] trained case-specific Random Forests (RF) to make mortality prediction and exhibited the best AUROC compared with other single models such as death counting, logistic regression, and decision tree.

No matter which method we use, feature selection is an important part. First, medical databases store vast amounts of clinical events and not all of them are related to the target task. Second, minimal clinical events enable doctors to make timely decisions. For severity scoring systems, a set of alternated related events is chosen based on clinicians' experience. A simple subset of those events is selected according to correlation coefficient or other index associated with the target concept of the prediction task [2–4,10]. Traditional machine learning and deep learning algorithms usually take the same features directly used in

* Corresponding author at: Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.
E-mail addresses: zhangfan@ict.ac.cn (F. Zhang), luochunjie@ict.ac.cn (C. Luo), lanchuanxin@ict.ac.cn (C. Lan), zhanjianfeng@ict.ac.cn (J. Zhan).

https://doi.org/10.1016/j.tbench.2021.100004
Received 6 August 2021; Received in revised form 11 October 2021; Accepted 20 October 2021
Available online 3 November 2021

Table 1
Comparison of benchmarking works.

		[7]	[6]	[9]	This work
Number of features	Smaller feature set	✓	✓	✓	✓
	Larger feature set		✓	✓	✓
Feature type	Non-time series		✓	✓	✓
	Time-series	✓	✓	✓	✓
Feature selection methods	Severity score	✓	✓	✓	✓
	Machine learning			✓	✓
	Evolutionary computing GA			✓	✓
Classifications methods	Traditional machine learning	✓	✓	✓	✓
	DNN	✓	✓		✓
Prediction tasks	Mortality	✓	✓	✓	✓
	Length of stay	✓	✓		✓
	Phenotyping	✓			
	ICD-9 code group	✓	✓		✓

severity scores [6,7] or take a similar method to select features [9]. In this paper, we do an exhaustive evaluation of various feature selection methods, and our main contributions are listed below.

(1) Benchmarking feature selection methods including traditional severity scores, machine learning-based feature selection methods, and evolutionary computing GA for three clinical prediction tasks.
(2) Compare feature subsets selected by GA for different classifiers. The results show that for the mortality prediction task, the features chosen by GA are universal for different classifiers.

The rest of this paper is organized as follows: in Section 2, we provide an overview of the related work; in Section 3, we describe the dataset and methods we employed; in Section 4, the benchmarking experiments and results are reported and discussed in detail; in Section 5, we summarize the paper.

2. Related work

First, we summarize the feature selection algorithms that are applied in the medical field. Then we discuss the existing benchmarks on healthcare datasets, especially for MIMIC-III. The comparison of benchmarks is listed in Table 1.

The first severity scores proposed such as APACHE [4], APACHE-II [11], and SAPS [12] selected features based on experience of medical experts. Further work usually used statistical methods to calculate correlation coefficient associated with target prediction task, such as [2,13,14]. Since publication, all of the methods have been continuously modified to improve the predictive performance [15]. A lot of work employed GA to select risk factors and predict in-hospital mortality [9,10,16–19]. In this work, we report an exhaustive set of benchmarking results of feature selection methods, including GA.

Public datasets such as MIMIC-III have promoted the benchmarking of models for clinical prediction tasks. Purushotham et al. [6] benchmarked deep learning models based on an extensive feature set and get high Area Under the Receiver Operating Characteristic Curve (AUROC) and Area under Precision–Recall Curve (AUPRC). The complete feature set we used is the same as the feature set C in [6], which contains 136 time-series features and five non-time series features. Harutyunyan et al. [7] first benchmarked four clinical prediction tasks and presented a multitask classifier. The most significant difference between us and previous works is that we benchmark feature selection algorithms, especially GA, instead of classification or regression algorithms. Krishnan et al. [9] proposed a GA-based model to make mortality prediction. We extend the benchmark to the other two prediction tasks and combine GA with DNN models to get higher AUROC and AUPRC. Johnson et al. [20] reproduced 28 published works for mortality prediction, and the results showed that it is a big challenge to reproduce other people's work without public code.

Table 2
Summary statics of cohort selection.

Data	Total
Admissions in the MIMIC-III (V1.4)	58,976
The first admissions	46,520
First admissions of adult patients	38,424
Patients died 24 h after the admissions	35,643

3. Materials and methods

3.1. Dataset preprocessing

MIMIC-III is developed by the Massachusetts Institute of Technology (MIT)'s Laboratory for Computational Physiology and contains around 60,000 intensive care unit admissions. MIMIC-III (v1.4) consists of 46,520 distinct patients and 58,976 admissions, from where we select 35,643 admissions for our experiments. We extracted data from 5 commonly used tables, namely inputevents, outputevents, chartevents, labevents, prescriptions tables. The statistics of cohort selection are tabulated in Table 2. We selected the patient cohort based on the following criteria:

- Only adult patients, whose age was >15 years at the time of ICU admission, were selected.
- Only the first admission was included for each patient. This decision uses the earliest available data to predict and ensure similar data selection compared to other related works.
- We only include the patients who died 24 h after the first admission.

Because the original data from MIMIC-III has erroneous records such as missing values, inconsistent units, etc., we clean data according to [6], which includes the following procedures: (a) Unify the units. (b) Select one valid record. For multiple records simultaneously, take the average values for numerical data and bring the first for categorical data. (c) Re-sample and fill-in the data. Time-series data is divided into hours and a forward–backward imputation is done to impute the missing values.

3.2. Prediction tasks

For benchmark, we select three clinical prediction tasks which are important in critical care research and are commonly studied by machine learning researchers. The first is in-hospital mortality which is important for doctors to take effective actions for patient care in Intensive Care Units (ICUs) [9]. The second is ICD-9 code group prediction, where we divide the ICD-9 codes into 20 groups according to [6] and treat it as a multi-classification problem. The third is length of stay prediction, which is to predict the hospital stay after admission.

Table 3
Genetic algorithm.

Genetic Algorithm
Input: Dataset D(X, Y) , classifier C and target number of features N
Output: Optimal N features X' for C
1. Randomly generate N features from X as X_0
2. Calculate fitness for X_0, which is AUC for D(X_0, Y) and C.
3. Set $X'=X_0$
4. **while** i ≤ 10000 **do**
5. Generate a new feature set X_i by randomly replacing one feature in X'
6. Calculate fitness for X_i
7. **if** X_i.fitness > X'.fitness:
8. $X'=X_i$
9. **if** X'.fitness ⩾ 1
10. **return** X'
11. **return** X'

3.3. Feature selection/extraction methods

We extract 136 time-series features and five non-time series features as our full feature set according to [6]. Those features are selected based on clinical significance and missing rate while containing all features used in severity scoring systems such as SAPS-II and SOFA. Based on this feature set, different feature selection methods are evaluated including GA based methods, scoring methods and machine learning methods.

GA is a metaheuristic inspired by the process of natural selection. It can be used to generate high-quality solutions to optimization and search problems by the process of mutation, crossover, and selection [21]. It is proved to be useful in feature selection as a wrapper feature selection technique [9]. Table 3 lists the GA procedure we used.

For scoring methods, we choose two popularly used severity scores, namely SAPS-II and SOFA. SAPS-II [2] is designed to predict the probability of hospital mortality. It can be calculated based on 17 variables which can be expand into 20 raw features of our complete feature set. SOFA [3] score can be calculated based on 6 variables which can be expand into 17 raw features of our complete feature set.

For machine learning methods, Principal Component Analysis (PCA) [22] and Recursive Feature Elimination (RFE) are chosen.

PCA is a widely used filter feature extraction technique, which projects the data to a new orthogonal space and then chooses a few of the essential features to achieve dimensionality reduction. RFE is a wrapper feature selection technique that selects features based on the accuracy of the subsequent classifiers.

3.4. Classification/regression methods

For machine learning we use three common commonly used algorithms: decision tree, Bayesian ridge regression, and logistic regression. For DNN we use three types of deep models, namely Feedforward Neural Networks (FNN), Recurrent Neural Networks (RNN), and Multimodal Deep Learning Model (MMDL) according to [6].

4. Benchmarking results

Based on the MIMIC-III dataset, we report the experimental results for three prediction tasks, which answer the following questions: (a) Can DNN models use relatively small feature subsets to perform as well as the full feature set? (b) Whether the subset of features selected by GA is universal for different classifiers of the same task?

4.1. Mortality prediction task evaluation

Tables 4, 5 show the results of mortality prediction task. Because PCA cannot handle time-series data, it is blank for RNN and MMDL results. We can observe that: (a) Deep learning-based prediction models perform better than traditional machine learning-based models and obtain around 2%–20% and 10%–30% improvement for AUROC and AUPRC, respectively. (b) GA performs better and obtains around 6%–18% and 10%–40% improvement over other methods for AUROC and AUPRC, respectively. (c) Compared with using all features (141 features), GA gets similar or even better results with only 20 features.

Fig. 1 shows the result of applying the features selected by GA-MMDL to other classifiers. We can observe that: (a) Although GA is a wrapper feature selection method, the features that GA-MMD chooses can also be used to other classifiers and achieve almost as good results as the GA combined with the specific classifier.

4.2. ICD-9 code prediction task evaluation

We divided the dataset into 20 classes according to [6] and treated it as a multi-classification task. However, because Bayes and LR in the package of scikit-learn do not support multi-classification tasks, we perform binary classification for these two algorithms and then calculate the average AUROC and AUPRC as the final results.

Tables 6, 7 show the results of icd-9 code group prediction task. We can observe that: (a) Deep learning prediction models perform better than traditional machine learning models and obtain around 9%–25% and 20%–35% improvement for AUROC and AUPRC, respectively. (b) GA performs better and obtains around 1%–10% improvement over other methods for both AUROC and AUPRC. (c) Compared with using all features (141 features), GA gets similar or even better results with only 20 features.

Fig. 2 shows the result of applying the features selected by GA-MMDL to other classifiers. We can observe that: (a) Only for deep learning models, the features that GA-MMDL chooses achieve similar results with the GA combined with the specific classifier. For machine learning classifiers, the features that GA-MMDL chooses do not perform well.

4.3. Length of stay prediction task evaluation

Table 8 shows the results of the length of the stay prediction task. We remove the LR algorithm since it is not capable of processing regression problems. We can observe that: (a) GA performs better than others and obtains around 6%–30% improvement over other methods in terms of MSE (in hours). (b) Compared with using all features (141 features), GA gets similar or even better MSE with only 20 features, save time and money.

Fig. 3 shows the result of applying the features selected by GA-MMDL to other regressors. We can observe that: (a) Only for the RNN model, the features GA-MMDL selects have good performance. For the other classifiers, the features that GA-MMDL chooses do not perform well.

4.4. Statistical significance tests of GA

From the above results, we can see that GA always performs better than other feature selection methods. We think this is because as the number of iterations (epochs) increases, GA can reach the local optimum. If there are enough iterations, GA can even reach the global optimum. The price of high precision is time overhead. However, this feature selection method only need to be trained once offline and in actual application doctors can quickly make a diagnosis with a few features.

To further check whether GA's improved performance is statistically significant compared with others we conducted statistical tests. The

Table 4
AUROC of in-hospital mortality prediction task.

Algorithm		Score method (Features)		Feature extraction/selection (Features)			All features (141)
		SAPS-II (20)	SOFA (17)	PCA (20)	RFE (20)	GA (20)	
ML	DT	0.6055 ± 0.0171	0.6780 ± 0.0052	0.5856 ± 0.0119	0.7009 ± 0.0066	**0.7657 ± 0.0085**	0.7631 ± 0.0119
	Bayes	0.8002 ± 0.0046	0.8018 ± 0.0011	0.7672 ± 0.0057	0.8166 ± 0.0085	0.9158 ± 0.0058	**0.9177 ± 0.0047**
	LR	0.5448 ± 0.0042	0.5570 ± 0.0043	0.5824 ± 0.0691	0.5581 ± 0.0044	0.7206 ± 0.0090	**0.7348 ± 0.0053**
DL	FNN	0.7945 ± 0.0059	0.7978 ± 0.0036	0.7863 ± 0.0067	0.8034 ± 0.0089	0.9207 ± 0.0026	**0.9263 ± 0.0032**
	RNN	0.8459 ± 0.0017	0.8651 ± 0.0037	–	0.8309 ± 0.0016	**0.9326 ± 0.0064**	0.9312 ± 0.0023
	MMDL	0.8532 ± 0.0033	0.8781 ± 0.0003	–	0.8282 ± 0.0057	**0.9376 ± 0.0036**	0.9345 ± 0.0029

Table 5
AUPRC of in-hospital mortality prediction task.

Algorithm		Score method (Features)		Feature extraction/selection (Features)			All features (141)
		SAPS-II (20)	SOFA (17)	PCA (20)	RFE (20)	GA (20)	
ML	DT	0.1992 ± 0.0121	0.3196 ± 0.0035	0.1756 ± 0.0080	0.3619 ± 0.0087	**0.4652 ± 0.0024**	0.4564 ± 0.0141
	Bayes	0.3374 ± 0.0158	0.3795 ± 0.0154	0.2706 ± 0.0040	0.3790 ± 0.0183	0.6109 ± 0.0271	**0.6333 ± 0.0096**
	LR	0.1549 ± 0.0041	0.1750 ± 0.0021	0.1498 ± 0.0146	0.1696 ± 0.0087	0.4301 ± 0.0088	**0.4320 ± 0.0025**
DL	FNN	0.3548 ± 0.0139	0.3871 ± 0.0073	0.2900 ± 0.0196	0.4010 ± 0.0076	0.7050 ± 0.0059	**0.7122 ± 0.0119**
	RNN	0.4299 ± 0.0116	0.5454 ± 0.0077	–	0.4396 ± 0.0101	**0.7486 ± 0.0189**	0.7283 ± 0.0074
	MMDL	0.4410 ± 0.0116	0.5687 ± 0.0070	–	0.4439 ± 0.0153	**0.7551 ± 0.0093**	0.7389 ± 0.0057

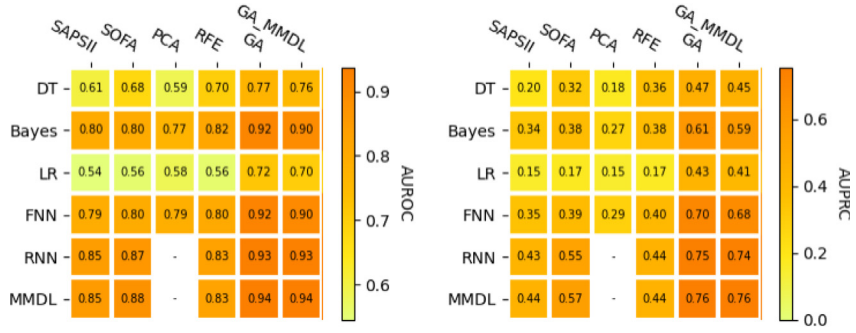

Fig. 1. Apply the features selected by GA-MMDL to other classifiers for the mortality prediction task.

Table 6
AUROC of icd-9 code group prediction task.

Algorithm		Score method (Features)		Feature extraction/selection (Features)			All features (141)
		SAPS-II (20)	SOFA (17)	PCA (20)	RFE (20)	GA (20)	
ML	DT	0.5735 ± 0.0004	0.5746 ± 0.0003	0.5609 ± 0.0005	0.5633 ± 0.0011	**0.5875 ± 0.0005**	0.5868 ± 0.0011
	Bayes	0.6818 ± 0.0010	0.6694 ± 0.0023	0.6523 ± 0.0027	0.6993 ± 0.0052	**0.7542 ± 0.0036**	0.7505 ± 0.0023
	LR	0.5454 ± 0.0004	0.5395 ± 0.0008	0.5438 ± 0.0006	0.5687 ± 0.0024	**0.6064 ± 0.0023**	0.6032 ± 0.0011
DL	FNN	0.8087 ± 0.0008	0.8034 ± 0.0014	0.8036 ± 0.0014	0.8158 ± 0.0002	0.8383 ± 0.0005	**0.8408 ± 0.0007**
	RNN	0.8147 ± 0.0012	0.8121 ± 0.0001	–	0.8229 ± 0.0003	0.8351 ± 0.0005	**0.8427 ± 0.0009**
	MMDL	0.8197 ± 0.0007	0.8179 ± 0.0003	–	0.8217 ± 0.0007	0.8384 ± 0.0007	**0.8440 ± 0.0007**

Table 7
AUPRC of icd-9 code group prediction task.

Algorithm		Score method (Features)		Feature extraction/selection (Features)			All features (141)
		SAPS-II (20)	SOFA (17)	PCA (20)	RFE (20)	GA (20)	
ML	DT	0.3719 ± 0.0012	0.3740 ± 0.0018	0.3522 ± 0.0013	0.3538 ± 0.0014	**0.3825 ± 0.0009**	0.3820 ± 0.0008
	Bayes	0.4523 ± 0.0006	0.4440 ± 0.0031	0.4096 ± 0.0031	0.4722 ± 0.0052	**0.5206 ± 0.0051**	0.5201 ± 0.0020
	LR	0.3440 ± 0.0004	0.3400 ± 0.0005	0.3414 ± 0.0002	0.3739 ± 0.0035	**0.4500 ± 0.0043**	0.3896 ± 0.0018
DL	FNN	0.6698 ± 0.0033	0.6602 ± 0.0022	0.6552 ± 0.0047	0.6717 ± 0.0003	0.7148 ± 0.0009	**0.7265 ± 0.0007**
	RNN	0.6820 ± 0.0006	0.6780 ± 0.0005	–	0.6897 ± 0.0008	0.7148 ± 0.0018	**0.7311 ± 0.0022**
	MMDL	0.6911 ± 0.0021	0.6902 ± 0.0004	–	0.6925 ± 0.0029	0.7214 ± 0.0015	**0.7340 ± 0.0005**

results are tabulated in Table 9. We can see that GA is statistically significant for mortality and length of stay prediction tasks but not for ICD-9 code group classification. This may be because it is more difficult to improve AUROC and AUPRC of multi-classification than binary-classification tasks.

5. Summary

This paper presented comprehensive benchmarking results of different feature selection methods and classification algorithms on three clinical prediction tasks. We demonstrated that: (a) GA always performs

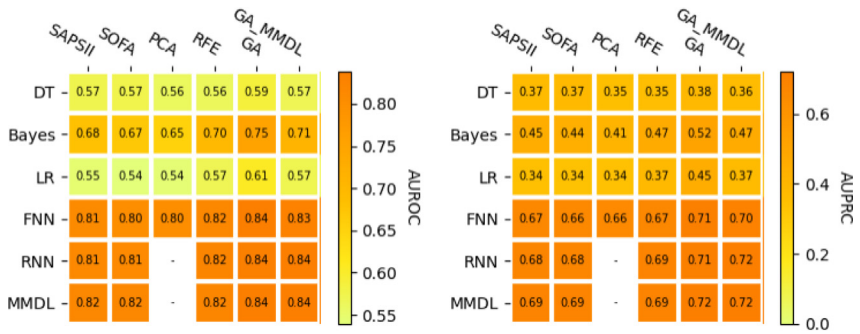

Fig. 2. Apply the features selected by GA-MMDL to other classifiers for icd-9 code prediction task.

Table 8
MSE of length of stay prediction.

Algorithm		Score method (Features)		Feature extraction/selection (Features)			All features (141)
		SAPS-II (20)	SOFA (17)	PCA (20)	RFE (20)	GA (20)	
ML	DT	54344.8815 ± 2390.9406	52836.9795 ± 3621.4408	54717.3409 ± 709.7653	48846.0888 ± 3862.0252	**43468.6583 ± 2223.4084**	45166.0388 ± 2853.9258
	Bayes	58715.9520 ± 2012.4531	74876.8679 ± 22148.1398	58784.5393 ± 3291.4266	61219.8744 ± 3515.9883	53172.5127 ± 4091.2336	**52564.0560 ± 2409.1296**
DL	FNN	60096.6133 ± 4174.1588	61843.7995 ± 6065.2225	57233.4701 ± 6296.6451	60358.6628 ± 4690.3527	**53843.4805 ± 3482.8460**	60998.5990 ± 9321.8411
	RNN	55031.8490 ± 3363.6808	54386.3086 ± 1304.0803	–	52148.4479 ± 1121.8866	43776.1263 ± 1633.5601	**42790.0521 ± 1290.5697**
	MMDL	54876.1159 ± 2560.8740	54138.1146 ± 2704.3223	–	51897.9544 ± 1752.6679	44385.0039 ± 3158.0891	**43398.7773 ± 4523.6935**

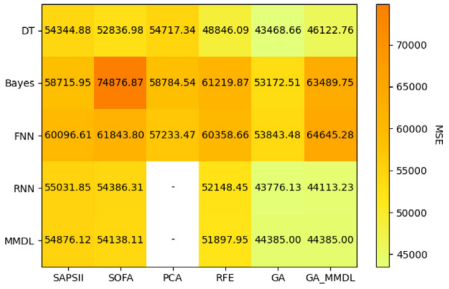

Fig. 3. Apply the features that GA-MMDL selects to other classifiers for the length of the stay prediction task.

Table 9
Whether is statistically significant with significance level 0.05.

Task	Metric	SAPS-II	SOFA	PCA	RFE
Mortality	AUROC	Yes	No	Yes	Yes
	AUPRC	Yes	Yes	Yes	Yes
ICD9	AUROC	No	No	No	No
	AUPRC	No	No	No	No
Length of stay	MSE	Yes	Yes	Yes	Yes

better than other feature selection methods; for mortality and length of stay tasks, the improved performance is statistically significant. (b) Compared with using all features, GA gets similar or even better predictive results with much fewer features, save time and money, which makes it more advantageous to detect and collect clinical data. (c) Other classifiers can also use the features that GA-MMDL selects for the mortality prediction task, achieving good performance.

As part of future work, we plan to make a severity-scoring system based on the features that GA-MMDL selects for the mortality task. This system promises doctors to quickly and accurately assess the severity of a patient's disease with a few simple variables.

References

[1] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[2] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, JAMA 270 (24) (1993) 2957–2963.

[3] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, L.G. Thijs, The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure, Springer-Verlag, 1996.

[4] G. Bhandoria, J.D. Mane, Can surgical apgar score (SAS) predict postoperative complications in patients undergoing gynecologic oncological surgery? Indian J. Surg. Oncol. 11 (1) (2020) 60–65.

[5] R. Pirracchio, M.L. Petersen, M. Carone, M.R. Rigon, S. Chevret, M.J. van der Laan, Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): a population-based study, Lancet Respir. Med. 3 (1) (2015) 42–52.

[6] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, J. Biomed. Inform. 83 (2018) 112–134.

[7] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (1) (2019) 1–18.

[8] J. Lee, Patient-specific predictive modeling using random forests: an observational study for the critically ill, JMIR Med. Inform. 5 (1) (2017) e3.

[9] G.S. Krishnan, S. Kamath, A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data, Appl. Soft Comput. 80 (2019) 525–533.

[10] A.E. Johnson, A.A. Kramer, G.D. Clifford, A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy, Crit. Care Med. 41 (7) (2013) 1711–1718.

[11] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, APACHE II: a severity of disease classification system, Crit. Care Med. 13 (10) (1985) 818–829.

[12] J.-R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, D. Villers, A simplified acute physiology score for ICU patients, Crit. Care Med. 12 (11) (1984) 975–977.

[13] M. Hoogendoorn, A. El Hassouni, K. Mok, M. Ghassemi, P. Szolovits, Prediction using patient comparison vs. modeling: a case study for mortality prediction, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2016, pp. 2464–2467.

[14] J. Calvert, Q. Mao, J.L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chettipally, R. Das, Using electronic health record collected clinical variables to predict medical intensive care unit mortality, Ann. Med. Surg. 11 (2016) 52–57.

[15] R. Pirracchio, Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sicula) project, in: Secondary Analysis of Electronic Health Records, Springer, 2016, pp. 295–313.

[16] R. Ahmad, P.A. Bath, Identification of risk factors for 15-year mortality among community-dwelling older people using cox regression and a genetic algorithm, J. Gerontol. (A Biol. Sci. Med. Sci.) 60 (8) (2005) 1052–1058.

[17] L.J. Adams, G. Bello, G.G. Dumancas, Development and application of a genetic algorithm for variable optimization and predictive modeling of five-year mortality using questionnaire data, Bioinform. Biol. Insights 9 (2015) BBI–S29469.

[18] C.-L. Chan, H.-W. Ting, Constructing a novel mortality prediction model with Bayes theorem and genetic algorithm, Expert Syst. Appl. 38 (7) (2011) 7924–7928.
[19] M.C. Engoren, R. Moreno, D.R. Miranda, A genetic algorithm to predict hospital mortality in an ICU population, Crit. Care Med. 27 (12) (1999) A52.
[20] A.E. Johnson, T.J. Pollard, R.G. Mark, Reproducibility in critical care: a mortality prediction case study, in: Machine Learning for Healthcare Conference, 2017, pp. 361–376.
[21] M. Mitchell, An Introduction to Genetic Algorithms, Vol. 1996, MIT press, Cambridge, Massachusetts. London, England, 1996.
[22] K. Pearson, LIII. on lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572.