AAAS
Health Data Science
Volume 2021, Article ID 9365125, 7 pages
https://doi.org/10.34133/2021/9365125

Health Data Science
A SCIENCE PARTNER JOURNAL

Research Article

Predicting Risk of Mortality in Pediatric ICU Based on Ensemble Step-Wise Feature Selection

Shenda Hong,[1,2] Xinlin Hou,[3] Jin Jing,[4,5] Wendong Ge,[4,5] and Luxia Zhang[1,2]

[1]*National Institute of Health Data Science at Peking University, Beijing, China*
[2]*Institute of Medical Technology, Health Science Center of Peking University, Beijing, China*
[3]*Neonatology Department of Peking University First Hospital, Beijing, China*
[4]*Harvard Medical School, Boston, MA, USA*
[5]*Clinical Data Animation Center (CDAC), Massachusetts General Hospital, Boston, MA, USA*

Correspondence should be addressed to Luxia Zhang; luxia_zhang@163.com

Received 15 November 2020; Accepted 21 January 2021; Published 16 June 2021

Background. Prediction of mortality risk in intensive care units (ICU) is an important task. Data-driven methods such as scoring systems, machine learning methods, and deep learning methods have been investigated for a long time. However, few data-driven methods are specially developed for pediatric ICU. In this paper, we aim to amend this gap—build a simple yet effective linear machine learning model from a number of hand-crafted features for mortality prediction in pediatric ICU. *Methods*. We use a recently released publicly available pediatric ICU dataset named pediatric intensive care (PIC) from Children's Hospital of Zhejiang University School of Medicine in China. Unlike previous sophisticated machine learning methods, we want our method to keep simple that can be easily understood by clinical staffs. Thus, an ensemble step-wise feature ranking and selection method is proposed to select a small subset of effective features from the entire feature set. A logistic regression classifier is built upon selected features for mortality prediction. *Results*. The final predictive linear model with 11 features achieves a 0.7531 ROC-AUC score on the hold-out test set, which is comparable with a logistic regression classifier using all 397 features (0.7610 ROC-AUC score) and is higher than the existing well known pediatric mortality risk scorer PRISM III (0.6895 ROC-AUC score). *Conclusions*. Our method improves feature ranking and selection by utilizing an ensemble method while keeping a simple linear form of the predictive model and therefore achieves better generalizability and performance on mortality prediction in pediatric ICU.

1. Introduction

Data-driven methods have been developed for mortality prediction in intensive care units (ICU) for a long time. Traditionally, table-based scoring systems such as Acute Physiology and Chronic Health Evaluation (APACHE) III score [1], Simplified Acute Physiology Score (SAPS) II [2], and Sequential Organ Failure Assessment (SOFA) [3] are more welcomed by clinical staffs because they are easy to calculate and understand. Recently, more sophisticated data-driven methods such as machine learning methods [4, 5], deep learning methods [4, 6–12], and ensemble methods [7, 9, 13] have been proposed for more accurate mortality prediction based on much larger public available ICU datasets such as Medical Information Mart for Intensive Care (MIMIC-III) [14] and eICU Collaborative Research Database (eICU) [15].

However, few data-driven methods are specially developed for pediatric ICU. For general ICU patients, some methods mixed the whole population without age stratification [4, 6, 8], others removed patients who are younger than an age threshold [5, 7, 13]. In fact, child patients in pediatric ICU have quite different physiological conditions compared with other patients. For example, hypernatremia is more severe for adult patients than neonate. So, we cannot bring general ICU models into pediatric ICU directly. Besides, although sophisticated machine learning methods usually get higher prediction accuracy, they are on the shelf because

clinical staffs cannot fully understand such black-box models [16]. The initial purpose of mortality prediction is to not play its role at all. In contrast, linear models and tree models are more interpretable and welcomed by clinical staffs—they can easily understand the predictions of simple models from mathematics.

In this paper, we aim to build a simple yet effective linear machine learning model from a number of hand-crafted features for mortality prediction in pediatric ICU. The model is trained based on a recent released pediatric ICU dataset named pediatric intensive care (PIC) database [17]. Unlike previous sophisticated machine learning methods such as random forest (RF) [18] classifier, gradient boosting machine (GBM) [19] classifier, or deep neural networks, we want our method to be kept simple that can be easily understood by clinical staffs. To achieve this, we first extract 397 hand-crafted features including demographics, routine vital signs, input and output, and laboratory values. Then, we propose an ensemble step-wise feature ranking and selection method to rank and select a small subset of effective features from the entire feature set. Finally, we build a logistic regression (LR) classifier upon selected features. The ensemble feature ranking and selection also promote the model generalizability so that it also performs well on the hold-out test set.

As a result, one of the final predictive linear model using 11 features achieves 0.7531 area under the receiver operating characteristic curve (ROC-AUC) score on hold-out test set, which is comparable with the linear model using all 397 features (0.7610 ROC-AUC score), and is higher than existing well known pediatric mortality risk scorer PRISM III (0.6895 ROC-AUC score). Another linear model using the best selected features (59 features) achieves a 0.7885 ROC-AUC score on the hold-out test set, which is the highest among all models. We release details of the models and hope they can be implemented for pediatric ICU mortality prediction in the real world.

2. Materials and Methods

2.1. Dataset. We use a recently released publicly available pediatric ICU dataset named PIC [17]. It comprises information relating to patients admitted to critical care units at the Children's Hospital of Zhejiang University School of Medicine in China. Existing data-driven analysis on PIC data include metabolic acidosis of children with acute kidney injury [20], timing/duration of tracheal intubation, and mechanical ventilation on mortality of children [21]. This paper will focus on a different task—mortality prediction.

We first exclude 191 patients whose mortality outcomes occurred within the first 24 hours following admission in the experiments. Then, we split the data into 80% development (for model training and validation) set, 20% test set (for evaluation). Mortality outcomes are determined by HOSPITAL_EXPIRE_FLAG in the ADMISSIONS table. Summaries of the patient characteristics of the development set and test set are presented in Table 1. The age distribution of the development set and test set is shown in Figure 1.

Table 1: Patient characteristics of development set and test set.

Demographic characteristics	Development set	Test set
# of subjects	10606	2652
Age in month (mean, std)	30.24, 44.35	28.96, 42.71
Gender		
Male (#, %)	6081, 42.66%	1535, 42.12%
Female (#, %)	4525, 57.44%	1117, 57.88%
In-hospital mortality (#, %)	630, 5.94%	150, 5.66%

2.2. Feature Extraction. We extract demographic information as features from patient's admission information and charted data. Besides, we also extract temporal values from routine vital signs, input and output, laboratory values, and transform them into features by taking min, max, and range values during the first 24 hours after admission. Features with more than 90% missing rates are removed. Consequently, the number of prepared hand-crafted features is 397.

2.3. Ensemble Step-Wise Feature Ranking and Selection. Feature selection can simplify the final model, which makes the predictive model easier understood and accepted by clinical staffs [22]. Feature selection can also promote the generalizability of the model, thus, leads to better performance on beyond development set (training data and validation data). Recent papers tried to promote the quality of feature selection by using ensemble methods [23, 24]. However, it is hard for many feature selection algorithms to control the number of output features, as they usually have other indirectly related hyperparameters.

Here, we propose an ensemble step-wise feature ranking and selection algorithm to control the number of output features in the development set. The framework is shown in Figure 2. The pseudocode is shown in Algorithm 1.

In the next sections, we will first introduce traditional feature ranking. Then, it leads to our ensemble feature ranking. Finally, we will show step-wise ensemble feature selection based on ranked features and build a machine learning model on selected features as the final mortality predictor.

2.3.1. Feature Ranking. We use GBM to rank feature importance. Each feature importance is calculated by the total reduction of the criterion from that feature, which is usually referred to as the average information gain (Gini importance) [19, 25].

Formally, denote the feature ranker as $\mathscr{H}$. The input is $\mathbf{x} \in \mathbb{R}^{n \times d}$, where n is the number of samples, and d is the number of features. Then, the output is $\mathbf{s} \in \mathbb{R}^{d}$, where $\mathbf{s}[i]$ represents the ith feature importance score.

$$\mathbf{s} = \mathscr{H}(\mathbf{x}). \tag{1}$$

2.3.2. Ensemble Feature Ranking. The ensemble method [26] can improve the performance of a single predictor, taking bagging [27] as an example. Formally, denote the ith predictor as $\mathscr{P}_i$, the input is $\mathbf{x}$ and output is $\mathbf{y}_i$. Then, we can get $\mathbf{y}_i = \mathscr{P}_i(\mathbf{x})$. Suppose we have M predictors, the ensemble output $\mathbf{y}$ based on bagging is calculated as $\mathbf{y} = 1/M \sum_{i=1}^{M} (\mathbf{y}_i)$.

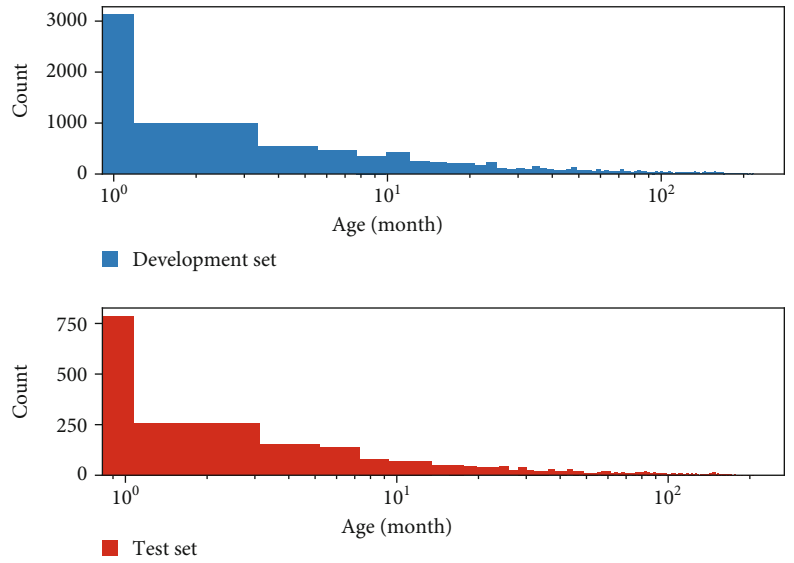

Figure 1: Age distribution of development set and test set (x-axes are log-scaled).

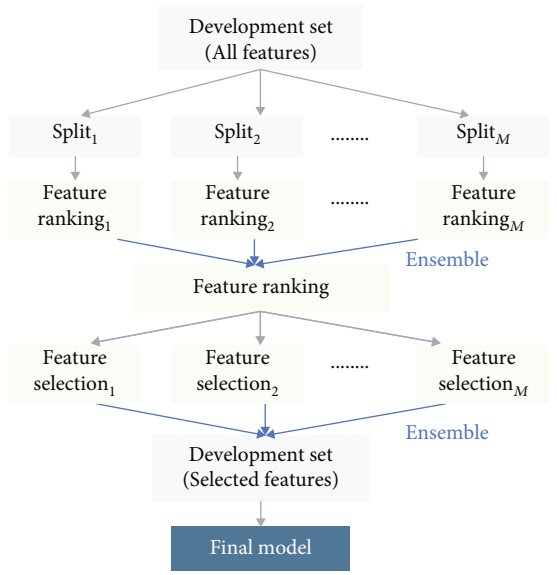

Figure 2: Framework of ensemble step-wise feature ranking and selection.

Given development set $\mathscr{D} = \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} \in \mathbb{R}^{n \times d}$, we split the entire dataset into M pieces of subsets, where 1 fold is validation set, and others are training set. The diagram is shown in Figure 3. Thus, we have M splits of training/validation set from the development set. Then, we can build M feature ranker based on them.

Here, we introduce bagging into feature ranking. Now, we have M feature ranker $\mathscr{H}_i$ where $i = \{1, 2, \cdots, M\}$, each feature ranker takes $\mathbf{x}_i$, a subset of $\mathbf{x}$, as input and output feature importance scores $\mathbf{s}_i$. In implementation, we will split the entire development set into pieces by cross-validation, then compose a subset $\mathbf{x}_i$ from them. Then, the ensemble feature score $\mathbf{s}$ is

$$\mathbf{s} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{s}_i) = \frac{1}{M}\sum_{i=1}^{M}(\mathscr{H}_i(\mathbf{x}_i)). \quad (2)$$

Finally, we will use $\mathbf{s}$ to rank features, which is more reliable than each single $\mathbf{s}_i$ based on ensemble theory.

2.3.3. Step-Wise Ensemble Feature Selection. Given $\mathbf{s}$, we can get the importance rank of features. Now, we will determine the number of select features based on it. This is done by step-wise ensemble feature selection. Intuitively, "step-wise" means we iteratively add one next feature to the current selected feature set; "ensemble" means we selected features by ensembling M predictors based on split shown in Figure 3.

In detail, at each iteration, we first select top j features based on feature ranker $\mathscr{H}$ to compose development set $\mathscr{D}_j = \{\mathbf{x}_j, \mathbf{y}_j\}$, where $\mathbf{x}_j \in \mathbb{R}^{n \times j}$. Then, we split $\mathscr{D}_j$ into M subsets. The ith training set is $\mathscr{D}_{ji} = \{\mathbf{x}_{ji}, \mathbf{y}_{ji}\}$, validation set is $\mathscr{T}_{ji} = \{\mathbf{x}_{ji}, \mathbf{y}_{ji}\}$. Next, we train mortality predictor

Input: Development set $\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{y}\}$, where $\mathbf{x} \in \mathbb{R}^{n \times d}$ is the data (n is the number of samples, d is the number of features), $\mathbf{y} \in \{0, 1\}^n$ is corresponding outcome.
Parameters: Maximum selected features K.
Output: Mortality predictor $\mathcal{F}$.
Split $\mathcal{D}$ into M subsets based on cross validation in Figure 3, the ith training set is $\mathcal{D}_i = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, validation set is $\mathcal{T}_i = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}$.
for i = {1,⋯,M} **do.**
    Build feature ranker $\mathcal{H}_i$ based on $\mathcal{D}_i$.
end for
Get ensemble feature ranker $\mathcal{H}$ from $\mathcal{H}_1, \cdots, \mathcal{H}_M$ based on Equation (2).
for j = {1,⋯,K} **do**
    Select top j features based on $\mathcal{H}$, compose development set $\mathcal{D}_j = \{\mathbf{x}_j, \mathbf{y}_j\}$, where $\mathbf{x}_j \in \mathbb{R}^{n \times j}$
    Split $\mathcal{D}_j$ into M subsets based on cross validation in Figure 3, the ith training set is $\mathcal{D}_{ji} = \{\mathbf{x}_{ji}, \mathbf{y}_{ji}\}$, validation set is $\mathcal{T}_{ji} = \{\boldsymbol{x}_{ji}, \boldsymbol{y}_{ji}\}$.
    for i = {1,⋯,M} **do**
        Training mortality predictor $\mathcal{F}_{ji}$, evaluate on validation set $\mathcal{T}_{ji}$
    end for
    Compute average performance of predictors with top i features
end for
Output top i features who has the best performance.

Algorithm 1: Ensemble step-wise feature ranking and selection.

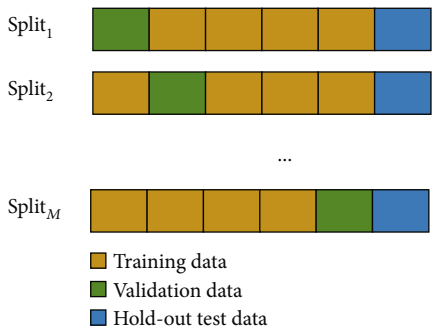

Figure 3: The diagram of training, validation, and test split.

$\mathcal{F}_{ji}$ based on $\mathcal{D}_{ji}$ and evaluate on validation set $\mathcal{T}_{ji}$. Finally, we compute the average performance of predictors who have top i features and output top i features who has the best performance.

2.4. Mortality Risk Predicting Model. To keep the final model simple and interpretable, we choose $\mathcal{F}$ to be logistic regression (LR) classifier. Hence, the predictors in step-wise feature selection are also LR classifier. To overcome the class imbalance problem, we use the inverse proportion of class frequencies to adjust sample-wise weights of the objective function.

Equation (3) gives the final model, where $\mathbf{x}$ is input features. Sigmoid is an activation function $\text{sigmoid}(x) = 1/1 + e^{-x}$. The details of each model can be found at https://github.com/hsd1503/PIC_mortality.

$$\mathbf{p} = \text{sigmoid}\left(\mathbf{w}^T\mathbf{x} + b\right). \tag{3}$$

2.5. Evaluations. We use the following measurements: receiver operating characteristic (ROC) curves, area under the ROC curves (ROC-AUC), precision-recall (PR) curves, and area under the PR curves (PR-AUC). For the development set, we further split it into 10 folds, where 1 fold is the validation set, and others are the training set. We impute missing values with the population median. No scale of original features to keep interpretability.

3. Results

Figure 4 shows the importance score of all 397 features in descent order. The x-axis is features, and the y-axis is feature importance score. The curve decreases rapidly from over 0.06 to around 0.005, then approaching 0 at around 350 features.

Figure 5 shows the performance of predictors in the step-wise ensemble feature selection process. The x-axis is the number of features used in predictors, which is LR in this case. The y-axis is the average ROC-AUC score of predictors on the 10-fold validation set. We can see that the performance increases rapidly in the beginning and reaches platforms at 11 features and 22 features. We select minimal features by using the top 11 features, medium features by using the top 22 features, and best features by using the top 59 features.

Figure 6 shows ROC curves, ROC-AUC scores, PR curves, and PR-AUC scores of different models evaluated on the same hold-out test set. The compared methods are LR with minimal features, LR with medium features, LR with best features, LR with all features, and a well-known pediatric mortality risk scorer PRISM III [28]. We can see that LR best achieves the highest 0.7885 ROC-AUC score. LR minimal is comparable with LR all but it uses much less features. PRISM III achieves a 0.6895 ROC-AUC score.

4. Discussions

We extract 397 features from the PIC dataset in the beginning, but only a small fraction of features are truly useful to

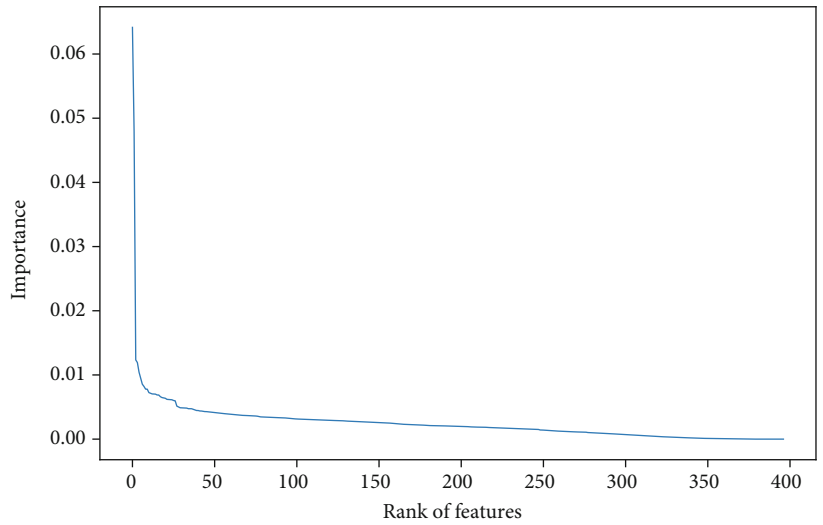

Figure 4: Importance score of features in descent order.

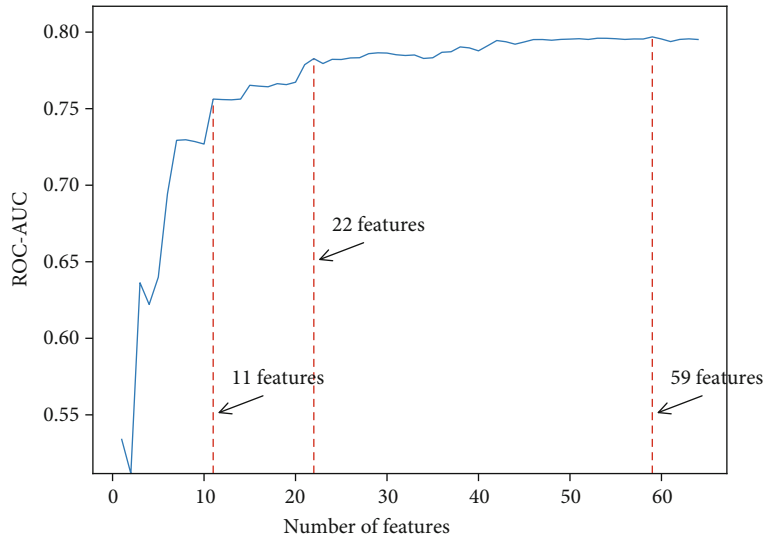

Figure 5: Performance of predictors in step-wise ensemble feature selection process.

predict mortality risk, while most features do not contribute too much. In Figure 4, a large majority of features are recognized as much less important than a few top features. This result is further verified by step-wise feature selection in Figure 5. The model achieves the highest performance when using the top 59 features, then dropping a little when adding more features. If all features are considered in the LR model, the ROC-AUC score is 0.7610 on the test set, which is only comparable with LR minimal.

Although some machine learning methods such as RF have the ability of feature selection when building the model, so that it can keep a good performance when modeling 397 features simultaneously. However, the final model is too sophisticated to be understood by humans, which hampers it to be implemented in the real world scenario. Besides, the high complexity of the model might also sacrifice the model generalizability.

Our ensemble feature selection has several advantages. First, feature ranking and feature selection are two steps to avoid overfitting, while such phenomena might happens in RF or (Least Absolute Shrinkage and Selection Operator) LASSO, as they simultaneously ranking, selecting, and build model. Second, use an ensemble of multiple folds further improving the generalizability of trained models. Third, the step-wise strategy can be used to control the number of features if special requirements are needed.

Our study has limitations. First, the dataset was collected from a single-center pediatric-specific hospital in China. Although it is a tertiary hospital and covers a broad area of East China, the population might not fully present Chinese

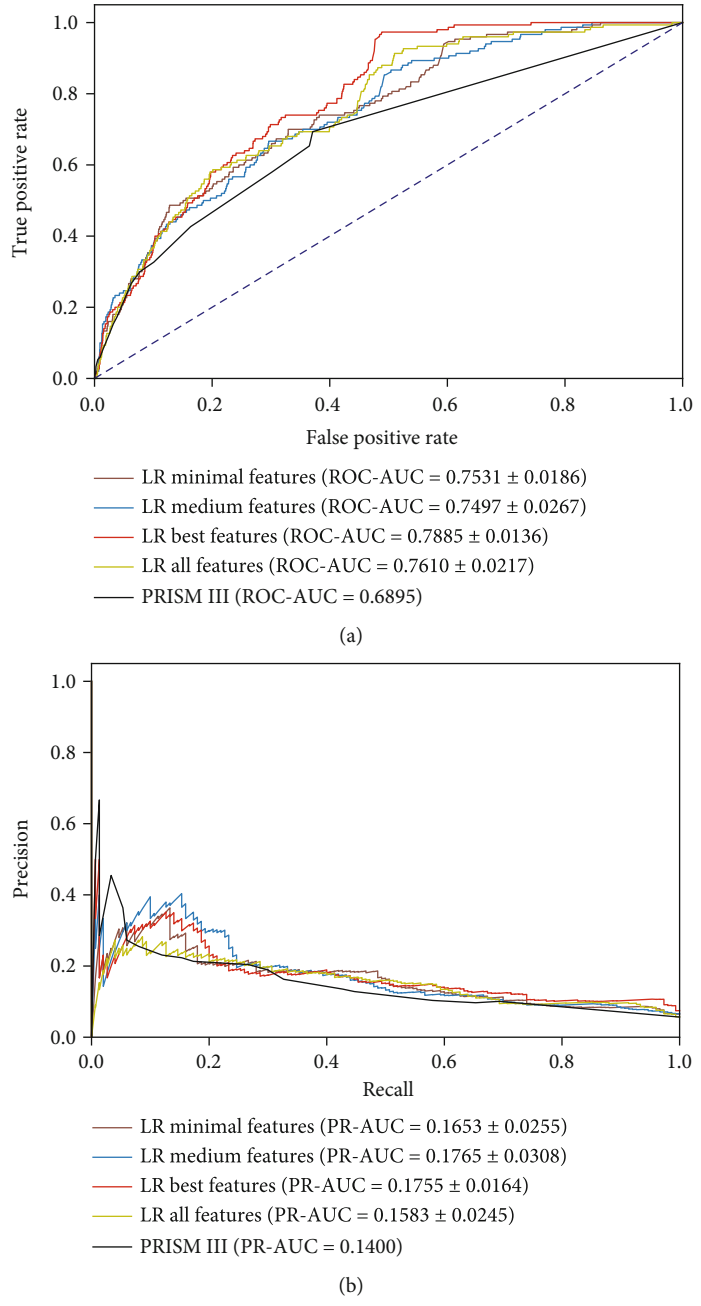

Figure 6: (a) ROC curves, average, and standard deviation ROC-AUC scores of different models. (b) PR curves, average scores, and standard deviation PR-AUC scores of different models.

children's demographics. Second, many features have high missing values ratios due to relatively low quality in the ICU dataset. In this paper, we impute missing values with population median. It might lead to inevitable biases in the model. Third, we did not perform clinical experiments in this paper. It will become our main direction for future work.

5. Conclusion

In this paper, we present the first work on the PIC dataset that build an interpretable machine learning model for mortality prediction in pediatric ICU. We propose an ensemble step-wise feature selection method to select a small subset of effective features from 397 features and then build a simple linear model for prediction.

In the future, we consider several ways for deployment in real-world applications. First, we will integrate this model into the existing Health Information System (HIS) and let the computer calculate. Second, we plan to simplify the LR model to be a table-based scoring system, so that clinical staffs can calculate it by hand. Moreover, we also plan to adjust weights and reduce features for better-customized deployment.

Conflicts of Interest

The authors have no competing interests to declare.

Authors' Contributions

Shenda Hong implemented the method and conducted the experiments. Shenda Hong and Xinlin Hou analyzed the results. All authors were involved in developing the ideas, discussing the results, and writing the paper.

References

[1] W. A. Knaus, D. P. Wagner, E. A. Draper et al., "The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, 1991.

[2] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new Simplified acute physiology score (saps ii) based on a European/North American Multicenter Study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.

[3] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J.-L. Vincent, "Serial evaluation of the sofa score to predict outcome in critically ill patients," *JAMA*, vol. 286, no. 14, pp. 1754–1758, 2001.

[4] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, 2019.

[5] K. Lin, Y. Hu, and G. Kong, "Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model," *International Journal of Medical Informatics*, vol. 125, pp. 55–61, 2019.

[6] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of Biomedical Informatics*, vol. 83, pp. 112–134, 2018.

[7] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "Raim: Recurrent attentive and intensive model of multimodal patient monitoring data," in *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2565–2573, New York, NY, USA, July 2018.

[8] J. Gao, C. Xiao, L. M. Glass, and J. Sun, "Dr. Agent: clinical predictive model via mimicked second opinions," *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1084–1091, 2020.

[9] S. Hong, Y. Xu, A. Khare et al., "Holmes: health online model ensemble serving for deep learning models in intensive care units," in *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1614–1624, New York, NY, USA, August 2020.

[10] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[11] Y. Si, J. Du, Z. Li et al., "Deep representation learning of patient data from electronic health records (EHR): a systematic review," *Journal of Biomedical Informatics*, article, 103671, 2020, https://www.sciencedirect.com/science/article/abs/pii/S1532046420302999.

[12] C. Sun, S. Hong, M. Song, and H. Li, "A review of deep learning methods for irregularly sampled medical time series data," 2020, http://arxiv.org/abs/2010.12493.

[13] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach," *International Journal of Medical Informatics*, vol. 108, pp. 185–195, 2017.

[14] A. E. Johnson, T. J. Pollard, L. Shen et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[15] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The EICU collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, no. 1, 2018.

[16] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[17] X. Zeng, G. Yu, Y. Lu et al., "Pic, a paediatricspecific intensive care database," *Scientific Data*, vol. 7, no. 1, p. 14, 2020.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[20] H. Morooka, D. Kasugai, A. Tanaka, M. Ozaki, A. Numaguchi, and S. Maruyama, "Prognostic impact of parameters of metabolic acidosis in critically ill children with acute kidney injury: a retrospective observational analysis using the pic database," *Diagnostics*, vol. 10, no. 11, p. 937, 2020.

[21] S.-L. Chong, T. K. Dang, T. F. Loh et al., "Timing of tracheal intubation on mortality and duration of mechanical ventilation in critically ill children: a propensity score analysis," *Pediatric Pulmonology*, vol. 55, no. 11, pp. 3126–3133, 2020.

[22] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[23] X. Song, L. R. Waitman, Y. Hu, A. S. L. Yu, D. Robins, and M. Liu, "Robust clinical marker identification for diabetic kidney disease with ensemble feature selection," *Journal of the American Medical Informatics Association*, vol. 26, no. 3, pp. 242–253, 2019.

[24] K. De Silva, D. Jönsson, and R. T. Demmer, "A combined strategy of feature selection and machine learning to identify predictors of prediabetes," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 39–406, 2020.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, 2012.

[27] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[28] M. M. Pollack, K. M. Patel, and U. E. Ruttimann, "PRISM III: an updated pediatric risk of mortality score," *Critical Care Medicine*, vol. 24, no. 5, pp. 743–752, 1996.