AN INTERPRETABILITY APPROACH FOR MORTALITY RISK PREDICTION BASED ON W-BDA AND MLP

Guanghua ZHANG[1], Huimin ZHANG[2], Mingxing FANG[3], Qi ZHANG[4], Renshuang DING[5,*]

Given the characteristics of acute respiratory distress syndrome (ARDS) medical data with unbalance, small samples and large feature space, and the lack of interpretability of existing model, this paper proposes an interpretable method for mortality risk prediction based on weighted balanced distribution adaptation (W-BDA) and multilayer perceptron (MLP). Firstly, the extracted ARDS data were preprocessed for the divided source and target domains. Secondly, feature selection based on XGBoost was performed in two domains to eliminate redundant features and achieve dimension reduction. Thirdly, the reconstructed domains were mapped to the same reproducing kernel Hilbert space (RKHS) through W-BDA, and the balance factor was introduced to achieve the weighted equilibrium adaptation of conditional and marginal distributions. Finally, the MLP network model was trained by the new source domain, and the mortality risk prediction of ARDS was achieved on the new target domain through parameter tuning and cross-validation. The experimental results show that the area under the receiver operating curve (AUC) of the method proposed in this paper is as high as 0.905 when predicting the risk of death, the accuracy is 87.78%. Compared with traditional methods, this method combined with SHAP could obtain better accuracy and reliable interpretability, providing more exact diagnosis advice for medical workers.

Keywords: W-BDA, MLP, mortality risk prediction, XGBoost, transfer learning

1. Introduction

Machine learning aiding medical diagnosis shows the excellent prospects [1], which have potential applications across multiple fields. ARDS is non-hydrostatic pulmonary edema associated with various etiologies defined by a common set of clinical features [2], with a high in-hospital mortality rate of

[1] Prof., School of Information Science and Engineering, Hebei University of Science and Technology, China, e-mail: xian_software@163.com.
[2] Master's degree candidate, School of Information Science and Engineering, Hebei University of Science and Technology, China, e-mail: zhm970228@163.com.
[3] Prof., Department of Critical Care Medicine, the Third Hospital of Hebei Medical University, China, e-mail: m18533112886@163.com.
[4] Master's degree candidate, Department of Critical Care Medicine, the Third Hospital of Hebei Medical University, China, e-mail: zq15200185008@163.com.
[5] * Lecturer, Corresponding author, School of Information Science and Engineering, Hebei University of Science and Technology, China, e-mail: abc618382@126.com.

approximately 10%~40% [3, 4]. The clinical prediction model established by machine learning technology can provide decision support for doctors to assess the condition and determine the treatment plan [5] and save more time for patients with early intervention for careful examination and treatment to reduce risk. However, most existing models applied in the intensive care unit (ICU) provide limited prognostic information and poor predictive power, thus causing controversy [4]. Jing et al. [6] established a risk prediction model for ARDS patients through a traditional logistic regression algorithm to assist patients with different risk stratification. Still, logistic regression cannot handle the correlation between features well, and the model was based on a few samples. Huang et al. [4] adopted the random forest algorithm, better than the existing scoring system based on logistic regression, but they could not make predictions beyond the data range of the training set, which may lead to overfitting in modeling some specific noisy data. Aktar et al. [7] compared various supervised machine learning algorithms suitable for clinical use, but the prediction performance depended on the size of the sample size, which was not conducive to promoting the model. Few single disease data and ample feature space are resulting in significant prediction errors, overfitting, and instability of the model. Therefore, it is still challenging to predict mortality of ARDS more accurately, efficiently, and reliably.

The clinical manifestations of some patients with specific diseases similar to ARDS are not typical, and the study sample size is small. At present, the primary means to solve the problem of small samples are transfer learning and sampling techniques [8]. Sampling technology generates balanced samples on the original samples based on a particular strategy and independently trains different prediction models of disease mortality risk, ignoring the knowledge transfer between other models [9]. As an essential branch of machine learning, transfer learning has been gradually applied to various tasks such as medical image segmentation [1], disease prediction [10], and complication prediction [11] in the medical field. Transfer learning breaks the assumption of independent and identical distribution in traditional machine learning [12]. By utilizing a small amount of labeled samples, cross-domain learning can be realized, and unmarked samples can be labeled [13]. For feature-based method, RKHS uses the maximum mean difference (MMD) as the metric [13] to minimize the data distribution difference between the two domains and perform dimensionality reduction processing, which transforms the problem into a transfer matrix learning problem and simplifies the optimization process [14]. Pan et al. [15] proposed TCA that attempted to learn a set of common transfer components between the source and target domains so that when the features of the two domains were mapped onto the common feature subspace, the difference in the data distribution of different domains could be significantly reduced. But the principal reduction of TCA is the difference in marginal distribution, and the contrast of conditional distribution is

not significantly reduced. Wang et al. [16] proposed BDA to adjust the difference between marginal and conditional distributions by introducing a balance factor. For the class imbalance in transfer learning, the W-BDA was also proposed, using class prior to accurately approximate the conditional distribution of the target class and training a classifier to ameliorate the performance, which solved the problem of slight sample imbalance to a certain extent.

Additionally, the existing machine learning model is equivalent to a black box in the prediction process [17]. The transfer learning method reduces the difference between the two domains. Although the prediction accuracy of a model is improved, it lacks interpretability. In the medical field, the risk of misdiagnosis is too significant, and it is not enough for practitioners to understand the accuracy of a model; but also need to know the characteristic basis of the model prediction [15]. Randomly selecting a subset of samples from any domain will increase the spatial difference of representation [13]. Therefore, unnecessary features are excluded for the two domains and provide a certain degree of interpretability so that medical staff can trust the model and its prediction results, which is convenient for promoting the model.

To sum up, it is difficult to train an efficient and exact mortality risk prediction model because of small samples, large feature space, imbalance of positive and negative samples in ARDS, and the existing models' lack of explicability. Therefore, an explicability method for mortality risk prediction based on W-BDA and MLP is proposed by adjusting the participants to improve the model prediction accuracy.

2. An interpretability method based on W-BDA and MLP

This paper proposes an interpretable mortality risk prediction method and its research framework is shown in Fig. 1. First, this study needs to extract ARDS data from the database and and other preprocessing operations. Since the W-BDA requires the source and target domains to be similar and isomorphic, we need to divide the extracted ARDS data into the two domains. Then, XGBoost is used to reduce the dimension between the domains, excluding the redundant features. The selected features form the new source (target) domain, and SHAP explains the features used for subsequent transfer learning. Interpreting the results independent of the predictive model used ensures the reliability of the results and provides more evidence support for solving clinical problems. And then, using MMD as the metric, the features of the new source (target) domains are mapped to the RKHS. The weighted equilibrium adaptation of the marginal and conditional distributions is realized in this space. Finally, the MLP model is used for training in the new source domain, which is tested on 30% of sample data in the new target domain through hyperparameter optimization, parameter adjustment, and cross-validation

later. The parameters with the best model performance are retained and compared with various model evaluation indicators and machine learning methods.

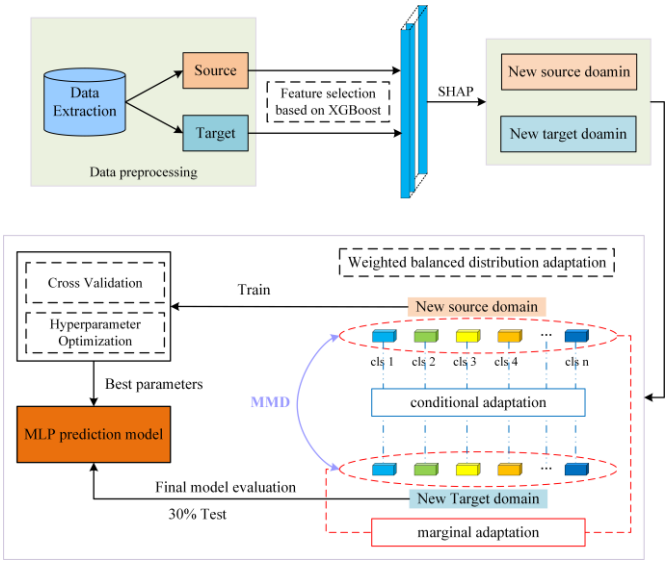


Fig. 1. Research framework of mortality risk prediction

2.1. Domain division and data preprocessing

This study investigated the mortality risk in adult ARDS patients and required data extraction from the MIMIC-III database. Since there are very few patients marked as ARDS in this database, which is not enough to carry out experimental research, this paper proposes the selection and inclusion criteria of ARDS patients and the data extraction process in combination with the Berlin definition criteria [18] and doctor's recommendations, as shown in Fig. 2. The patient selection process operates primarily in the following medical data tables and views: ADMISSIONS, PATIENTS, NOTEEVENTS, VENTFIRSTDAY, ICUSTAYS, CHARTEVENTS and BLOODGASFIRSTDAYARTERIAL. This database recorded the age of patients in their 90s as 300 years old, so we limited the age of patients to be less than 100 to avoid outliers. According to the clinical experience of doctors, the hospitalization time of ARDS patients generally does not exceed half a month, so this paper sets the hospitalization time of patients to 2 to 15 days. In addition, patients also need to meet the conditions of being admitted to the ICU for the first time, the minimum oxygenation index on the first day is less than 300, and have undergone chest imaging examination and mechanical ventilation, to more accurately screen as many ARDS patients as possible. The indicators in the reference view represent the extraction of the relevant indicators under study. For some indicators not in the view, this paper uses the label of the d_items table and the corresponding itemid to query the noun similarity, such as PaO_2, NBPM, PIP, PLAP, MAP, C.O., SaO_2 and SVR. After obtaining the itemid

corresponding to each indicator, we extract data from the CHARTEVENTS table, and the remaining indicators can be extracted from the related views. This paper removed patients with more than 30% missing data and pulled 4010 patients and 36 indicators, including demographics, ICU conditions, ventilator parameters, clinical indicators, and laboratory measurement information.

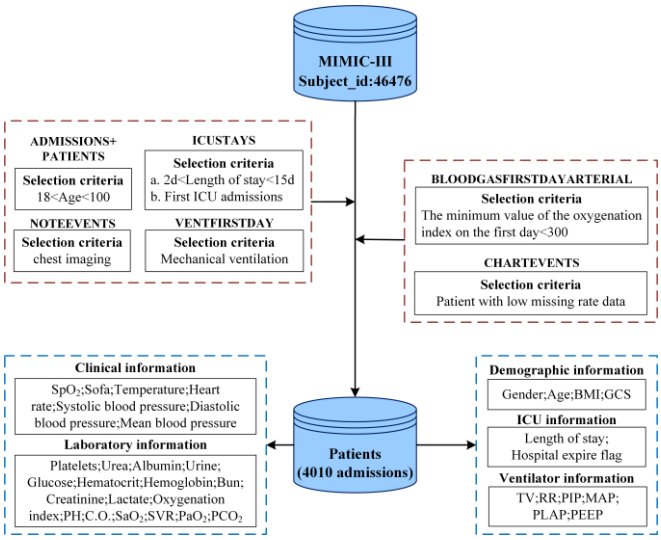


Fig. 2. Inclusion criteria for patient cohort

This study proposes a division method to divide the above extracted ARDS data into the source (target) domain to make the two domains similar and isomorphic. Firstly, the six indicators of gender, age, hospital stay, GCS, Sofa score and albumin are used as the common features of the feature space, in which gender needs to be converted into numerical features; that is, F is converted to 0, and M is converted to 1. If GCS, Sofa, and albumin have multiple records for a patient in the view, they are grouped by subject_id and averaged so that there is one record for one index for one patient. Then, according to the above ITEMID, each index measured by each patient during their stay in the ICU is averaged by day, and the maximum and minimum values of the average of the indexes measured every day are used as two different characteristics of the patient, to avoid the error caused by accidental abnormalities. After that, the maximum and minimum values of the remaining indicators included in the view are used as two different features. If there are multiple records of the same patient, these records are grouped by subject_id and averaged. Finally, the maximum and minimum values of each monitoring index extracted above are scrambled and randomly extracted. The extracted one is conducted as the source feature, and the remaining one is automatically classified into the target feature. The random principle ensures that each feature has the same possibility of being selected, avoiding the influence of subjective factors. This division method makes the source and target

domains have intersecting features. The values are distributed at different time nodes, so the feature space is similar but somewhat different. The feature division process of the two domains is shown in Fig. 3, and eventually, the source domain features (D_s) and the target domain features (D_t) are all obtained.

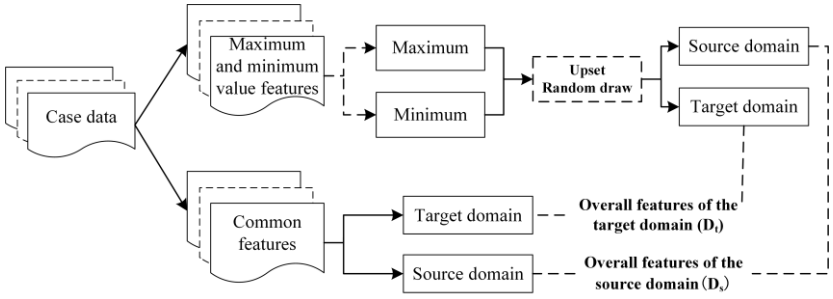


Fig. 3. Feature division process of source and target domains

The hospital expire flag is taken as the sample label, combined with $\{D_s, D_t\}$ as the source domain data and target domain data. At this time, there were some missing data in the two domains. We used the features corresponding to complete and non-missing clinical data points to retain more valuable data as a sample input. The missing values' features were used as k-nearest neighbor interpolation labels. Let $k = 5$, the missing values of incomplete samples are obtained from the adjacent five samples, so this interpolation method will not add too much redundant information.

2.2. SHAP interpretation based on XGBoost feature selection

XGBoost uses the 2nd derivative to calculate the objective function in the model optimization process while adding a regularization term to the objective function [19], which ensures high solution efficiency and increases scalability. The generalization ability of the algorithm can be effectively improved by sampling all the features of D_s and D_t respectively and training the feature combination. Then, the objective function of XGBoost to extract the patient's numerical features is expanded by Taylor, as shown in equation (1).

$$obj^{(t)} = \sum_{i=1}^{T}\left[\sum_{i\in I_t} g_i + \frac{1}{2}\left(\sum_{i\in I_t} h_i + \lambda\right)\right] + \gamma T + \frac{1}{2}\|w\|^2 \tag{1}$$

where, g_i in the first term represents the first-order partial derivative, h_i represents the second-order partial derivative; the remaining term is the regularization term, which is used to avoid overfitting during training. Then the data points are substituted into equation (1) for gradient descent to seek the feature combination of the optimal solution.

Based on game theory and local interpretation, the core of SHAP is the

Shapely value, a method to describe the contribution of each feature when a model predicts a specific data point [20]. It facilitates the clinician in analyzing the reliability of the model prediction. In this paper, feature selection was performed on the features of the two domains respectively. Then, the Shapley value of each index was calculated to analyze the correlation between each feature and mortality risk. Finally, the selected important features were explained in combination with relevant clinical research results and clinical manifestations.

2.3. W-BDA algorithm

While adapting to the conditional and marginal distributions between domains, BDA could also exploit the importance of these two distributions to adapt to specific transfer learning tasks effectively [16]. Aiming at the imbalance of ARDS sample categories, this paper introduces W-BDA, an extended BDA algorithm. It could also adaptively adjust the weight of each category while considering the distribution adaptability between domains. Taking MMD as the metric, this paper maps X_s and X_t formed by the initial selection of XGBoost into the same RKHS to estimate the difference between the two distributions. The definition of MMD is shown in equation (2).

$$D(X_s, X_t) = \mu \sum_{c=1}^{C} \left\| \frac{1}{n_c} \sum_{x_{s_i} \in X_s^{(c)}} x_{s_i} - \frac{1}{m_c} \sum_{x_{t_j} \in X_t^{(c)}} x_{t_j} \right\|_H^2 + (1-\mu) \left\| \frac{1}{n} \sum_{i=1}^{n} x_{s_i} - \frac{1}{m} \sum_{j=1}^{m} x_{t_j} \right\|_H^2 \quad (2)$$

where, the 1st term represents the conditional distribution distance between the two domains, while the 2nd term represents the marginal one; H stands for RKHS. $0 \le \mu \le 1$, when μ approaches 0, it means that the two domains are not similar, so the marginal distribution dominates; when μ approaches 1, it means that the two domains are similar, so the adaptation of the conditional distribution needs to be focused. Therefore, the balance factor μ could adaptively adjust the importance of each distribution to produce fine results.

The smaller the value of MMD, the smaller the difference between the distribution of two reconstructed domains. This paper introduces the kernel matrix $K = \psi(X)^{\cdot} \psi(X) \in \mathrm{R}^{(n+m)\times(n+m)}$, further exploiting matrix trace and regularization, and optimizes the equation (2) by the Lagrange multiplier $\Phi = diag(\phi_1, \cdots, \phi_d)$, which can be transformed into equation (3).

$$\left(K \left(\mu \sum_{c=1}^{C} W_c + (1-\mu) M_0 \right) K + \lambda I \right) A = KHKA\Phi \quad (3)$$

where, s.t. $\mathrm{A}^{\cdot} XH^{\cdot} A = I$, $0 \le \mu \le 1$ are constraints; λ is the regularization parameter; A represents the transformation matrix; $I \in \mathrm{R}^{(n+m)\times(n+m)}$ represents the

identity matrix and H represents the center matrix, $H = I - (1/n)I$. M_0 and the weight-optimized W_c are constructed by MMD matrix.

Therefore, the optimization problem is transformed into a generalized eigendecomposition problem, and the optimal transformation matrix A is obtained by solving equation (3) to obtain the first d minimum eigenvectors. In this paper, $d = 20$ was set to represent the number of dimensions to be reduced, and the optimal migration result was obtained by searching for the balance factor.

2.4. MLP network prediction model and hyperparameter optimization

Basic MLP consists of an input, hidden, and output layer, a feedforward neural network that maps input data to a group of output data [22]. The neural network can learn the complex relationship between the features of the new source (target) domain after W-BDA mapping. The output description of MLP is shown in equation (4).

$$y_p = \varphi_o \left\{ \sum_{j=0}^{N} w_{jp}^{o} \left[\varphi_h \left(\sum_{i=0}^{M} w_{ij}^{h} x_i \right) \right] \right\} \tag{4}$$

where, x_i represents the input of i features of a given sample; $\{w_{ij}, w_{jp}\}$ indicates the weights between two layers; $\{\varphi_h, \varphi_o\}$ indicates activation functions.

As shown in Fig. 4, the MLP network architecture constructed in this paper includes two hidden layers with the ReLU activation function, where, the number of neurons in the input layer is the same as the number of input features; the number of neurons in the first hidden layer is twice that of the input layer; the second hidden layer has the same number of neurons as the input layer; the output layer has only one neuron, and its output is 0 or 1, representing the two outcomes of survival and death respectively.

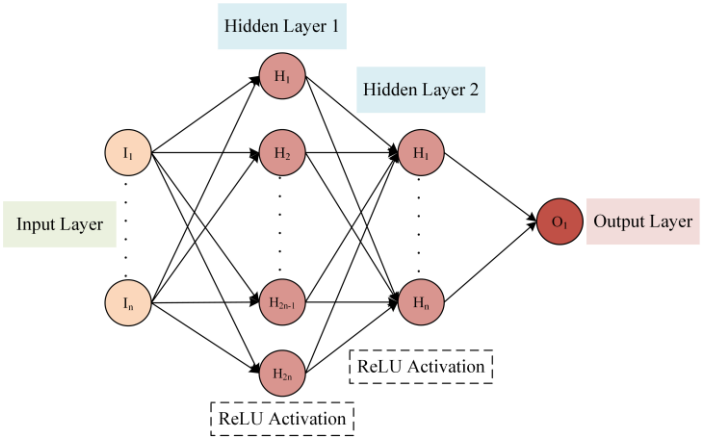


Fig. 4. The MLP neural network architecture

A SGD optimizer and a cross-entropy loss function are employed to compute gradients and weight updates after each input instance. The source domain sample features mapped by W-BDA are used as training data firstly input by the input layer. Then, through two fully connected layers, the neurons in each fully connected layer can fit the original data. Finally, the output layer outputs the data. It uses the output value and the new source domain sample label to construct the loss function, iteratively reducing the loss function through the gradient descent algorithm of back-propagation and updating the model parameters to minimize the value of the loss function. At this time, the MLP model can accurately fit the sample characteristics. For the regularization term parameters and learning rate of the model, this paper uses grid search cross-validation for combined search adjustment, and then by changing different hyperparameters, the model parameters with the highest performance are retained as the best model parameters.

In the optimization process of each set of hyperparameters, this paper used five-fold cross-validation to ensure that each sample subset of the source domain after W-BDA mapping could be trained and capture features. At the same time, the AUC of each training was calculated, and the results of five curve fittings were averaged as the final AUC results of each training of the MLP training model. After all hyperparameter combinations were iterated, it selected the model parameter with the highest average AUC score to make predictions on 30% of the target domain samples after W-BDA mapping as test data.

3. Experiment and analysis

The operating system configuration of this experiment is Windows 10. The data is accessed by creating a PostgreSQL database connection, and with SQL programming, the ARDS data extraction process is implemented to obtain the experimental dataset. The division process of the source (target) domain and all subsequent experiments are implemented by Python 3.8 programming.

3.1. Experimental dataset

According to the division method of the source (target) domain proposed in this paper, the 36 indicators extracted related to the risk of ARDS death can derive 64 dimensional features. Therefore, 4010 inpatient records are composed of 64 features to form an experimental data set with large feature space and small sample size. The feature information of the dataset is shown in Table 1.

Table 1

Feature information of the experimental dataset

Data information	Source domain features (D_s)	Target domain features(D_t)
Demographics	gender, age, bmi_min, gcs	gender, age, bmi_max, gcs
ICU situation	icu_stay	icu_stay

Ventilator parameters	tv_min, resprate_min, pip_max, map_min, plap_max, peep_max	tv_max, resprate_max, pip_min, map_max, plap_min, peep_min
Clinical indicators	spo2_max, sofa, tem_min, hr_min, sysbp_max, diasbp_min, nbpm_max	spo2_min, sofa, tem_max, hr_max, sysbp_min, diasbp_max, nbpm_min
Laboratory measurement	platelets_min, urea_n_max, albumin, urine_max, glucose_max, lactate_max, hematocrit_min, pao2_max, hemoglobin_min, bun_max, creatinine_min, pco2_max, pao2fio2_max, ph_max, co_max, sao2_min, svr_max	platelets_max, urea_n_min, albumin, urine_min, glucose_min, lactate_min, hematocrit_max, pao2_min, hemoglobin_max, bun_min, creatinine_max, pco2_min, pao2fio2_min, ph_min, co_min, sao2_max, svr_min

After division, the feature dimensions of the two domains are the same, and the amount of data is the same. The two fields share the same characteristics, but also have some differences. In view of the obvious differences of medical data at different time nodes, this division of source and target domains has practical significance.

3.2. Evaluation indicators

Since the proportion of positive and negative samples in the dataset is quite different, to avoid the impact of unbalanced samples on the evaluation indicators, this experiment takes AUC as one of the evaluation indicators. The experiment also uses four evaluation indexes widely adopted in medical research, accuracy (Acc), precision (Pre), recall (Rec) and F1 score ($F1$), as shown in equation (5) ~ (6), to comprehensively evaluate the effectiveness of the method.

$$Acc = \frac{TN + TP}{TP + FP + FN + TN}; F1 = \frac{2TP}{2TP + FP + FN} \tag{5}$$

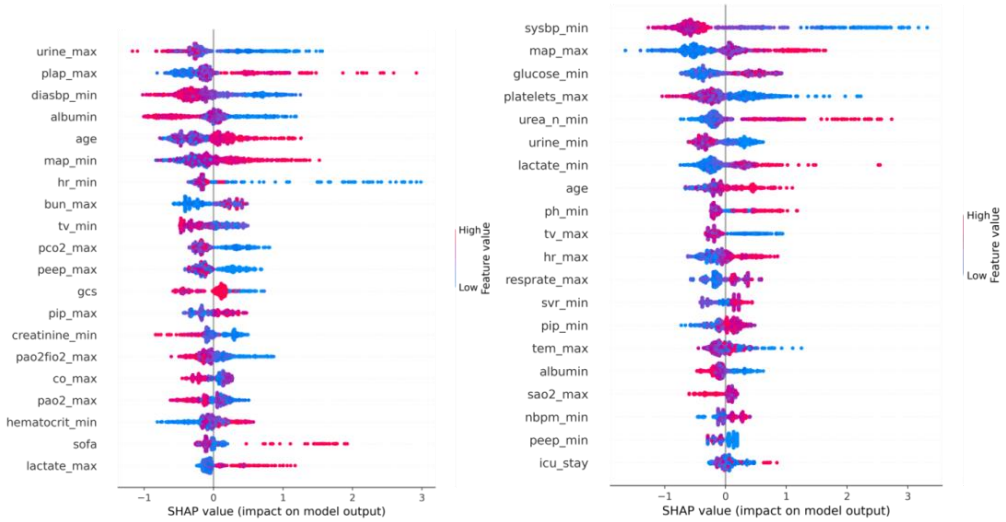
$$Pre = \frac{TP}{TP + FP}; Rec = \frac{TP}{TP + FN} \tag{6}$$

3.3. Experimental results and analysis

3.3.1. SHAP interpretability analysis

Combining the influence of features, the summary results of the SHAP model interpretation of XGBoost feature selection are shown in Fig. 5 (a) ~ (b), showing the top 20 clinical features that are highly correlated with ARDS mortality risk in the source and target domains, as well as the features after dimensionality reduction. According to the source domain results, urine_max played a crucial role in ARDS mortality risk prediction. Urine is one of the most important variables affecting renal function, and ARDS patients extracted in this paper may impact the kidneys. The lower the eigenvalue and the higher the

Shapley value, the higher the probability of developing ARDS mortality risk. Studies have shown that urine is associated with ARDS mortality [23], supporting our findings. PLAP is an essential indicator for ventilator detection of ARDS, which can effectively reflect the risk of barotrauma. Its maximum value (plap_max) is the second important feature, and the higher the feature value, the higher the probability of ARDS mortality risk. In addition, regarding laboratory tests, higher hematocrit_min values are associated with a higher risk of ARDS mortality but are also significantly less critical.



(a) Source domain (b) Target domain

Fig. 5. Source (target) domain SHAP summary graph

According to the target domain results, the minimum systolic blood pressure value (sysbp_min) contributed to ARDS mortality risk prediction. The MAP is the average pressure experienced by the lungs during the respiratory cycle, whose maximum value (map_max) is the second most important feature in the target domain. The higher the feature value, the higher the probability of mortality risk. This finding is in line with clinical results. Clinical measures are taken to increase expiratory resistance, appropriately expand MAP, and reduce the pressure difference between inside and outside the airway to prevent airway trapping and maintain patient exhalation [24]. In addition, the importance of the features of urine_min and urea_n_min is also apparent. Urea nitrogen is an essential indicator in biochemistry, and the higher its minimum characteristic value is, the more likely ARDS death will occur. To sum up, the importance of urine output to the two domains is ranked relatively high, indicating that this indicator has certain research significance for ARDS mortality risk prediction as a whole. The above experimental results are consistent with clinical and related

research results, further confirming the effects of urine, hematocrit, systolic blood pressure, MAP, urea nitrogen and related vital signs on ARDS mortality risk. It is important to note that the above summary graph illustrates the association between characteristics and ARDS mortality risk rather than causality. Therefore, it is essential to combine this information with the clinical experience of doctors and the condition of patients to determine whether this feature is an option for intervention.

3.3.2. Comparative analysis of performance of multiple methods

Next, this paper performs W-BDA mapping on the new source (target) domain composed of above 20 features. The MLP model constructed in this paper is trained by the new source domain and then randomly selects 30% of the samples from the new target domain for label prediction. Simultaneously, this research horizontally compares the prediction results of TCA, BDA, W-BDA and no transfer combined with decision tree (DT), Bayesian, 3NN, AdaBoost and MLP, respectively; longitudinally compares TCA, BDA, and W-BDA under each kernel function combining the accuracy, precision, recall, and F1 score of the DT, Bayesian, 3NN, AdaBoost and MLP. The results are shown in Table 2~5.

Table 2

Accuracy (%) of TCA, BDA, W-BDA and no transfer

Classification model	No transfer	Kernel	TCA	BDA	W-BDA
Decision tree	13.34	rbf	80.20	80.60	81.77
Bayesian	13.12	rbf	84.51	84.51	84.51
3NN	22.87	primal	80.78	81.67	82.04
AdaBoost	43.22	rbf	68.98	69.66	70.20
MLP	46.79	rbf	85.56	86.88	**87.78**

Table 3

Precision (%) of TCA, BDA, W-BDA and no transfer

Classification model	No transfer	Kernel	TCA	BDA	W-BDA
Decision tree	56.16	rbf	88.17	88.48	88.39
Bayesian	13.12	rbf	87.04	87.04	87.04
3NN	79.34	primal	87.76	88.23	**88.69**
AdaBoost	**95.96**	rbf	87.02	87.36	87.83
MLP	80.08	rbf	84.89	85.67	86.88

Table 4

Recall (%) of TCA, BDA, W-BDA and no transfer

Classification model	No transfer	Kernel	TCA	BDA	W-BDA
Decision tree	11.18	rbf	89.18	90.61	90.70

Bayesian	13.01	rbf	96.56	96.56	96.56
3NN	12.74	primal	89.57	89.89	90.93
AdaBoost	36.17	rbf	74.77	75.28	76.26
MLP	67.36	rbf	96.98	97.67	**98.05**

Table 5

F1 score (%) of TCA, BDA, W-BDA and no migration

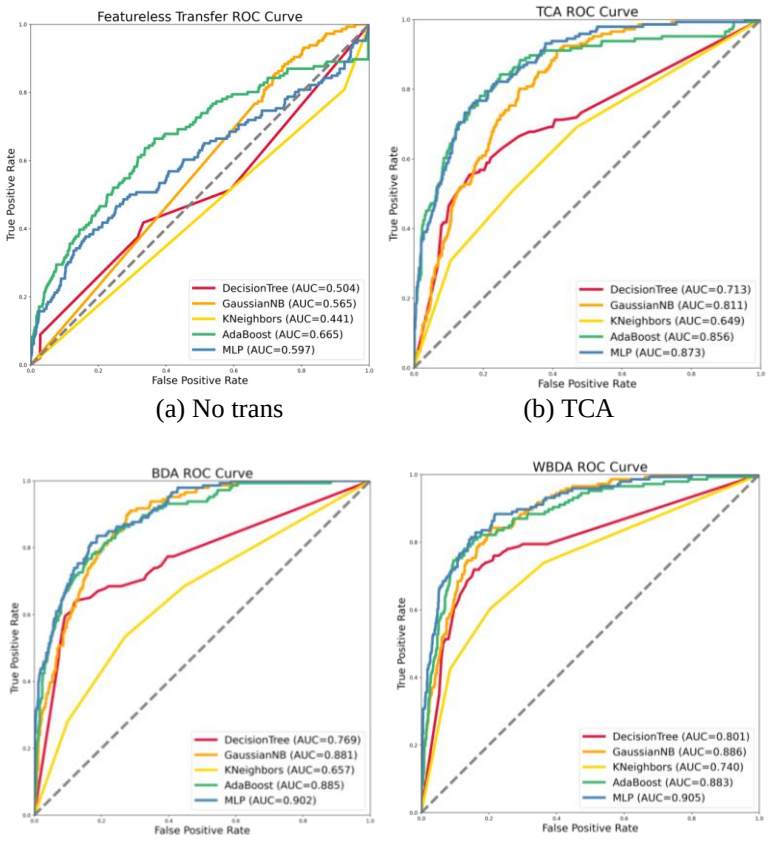
Classification model	No transfer	Kernel	TCA	BDA	W-BDA
Decision tree	18.65	rbf	88.67	89.53	89.53
Bayesian	13.06	rbf	91.55	91.55	91.55
3NN	21.95	primal	88.66	89.05	89.80
AdaBoost	52.54	rbf	80.43	80.87	81.64
MLP	73.17	rbf	90.53	91.28	**92.13**

In addition to the precision of AdaBoost, on the whole, the results of various evaluation indicators of the three transfer learning methods are much higher than those of the no-transfer method. We can see from Table 2 that the combination model of feature-based transfer learning and machine learning could improve the overall accuracy of the training by more than half. Compared with BDA and TCA, W-BDA has the highest accuracy when combined with any prediction model. We can learn from Table 4 and Table 5 that the comparison results of the recall and F1 score are generally consistent. The prediction effect of BDA is better than that of TCA, and W-BDA not only adjusts the weight of the distribution between domains for the new source and target domains but also performs class balance. Hence, the prediction effect is better than that of BDA. The model performance of the W-BDA and MLP has reached more than 90%, of which the F1 score reaches 92.13%. However, the precision results are opposite to those of the other three evaluation metrics. As shown in Table 3, the combination of W-BDA and 3NN has the highest precision. The combination of W-BDA and MLP is relatively poor, but the gap is insignificant. This is related to the kernel function adopted by transfer learning, and the combination of 3NN and Primal-based feature mapping method achieves the highest precision. In general, the interpretability method based on W-BDA and MLP for mortality risk prediction proposed in this paper performs the best in terms of precision, recall, and F1 score, whose overall performance is slightly higher than other methods.

3.3.3. AUC comparative analysis

To comprehensively assess the effectiveness of the proposed method due to the sample imbalance, this paper draws the ROC curve charts respectively. Five cross-validations were performed for each machine learning classification model. The AUC results of the five cross-validations were averaged as the final model

AUC and displayed on the ROC curve, as shown in Fig. 6 (a) ~ (d). Fig. 6 (a) reflects the AUC results of five models for two-domain features without any transfer, but the results presented are generally lower. The AUC obtained by AdaBoost training is the highest, but it is only 0.665; MLP is slightly lower than AdaBoost and cannot provide a reliable prediction effect. Some related studies like to use KNN as the baseline algorithm, but the performance achieved by model training is generally not high. The same is true in this paper. Among the five machine learning algorithms, the AUC trained by 3NN is the lowest, and the probability of being inferior to human judgment is high.



(a) No trans (b) TCA

(c) BDA (d) W-BDA

Fig. 6. AUC comparison chart

Fig. 6(b) ~ (d) respectively reflect that under the optimal kernel function, TCA, BDA, W-BDA and five models respectively combine the predicted AUC results. Among these methods, W-BDA maps the feature space so that the AUC obtained by the training of the classification model is the highest. Even the AUC of the mortality prediction model combined with BDA, W-BDA, and MLP respectively reaches more than 0.9, which is far higher than its combined effect

with DT and 3NN. Among them, the AUC achieved by the combined prediction of W-BDA and MLP is 0.003 higher than the combined prediction performance of BDA and MLP. Thus, the effectiveness of W-BDA for inter-domain class balance adjustment is illustrated. Through the vertical and horizontal comparisons, the W-BDA combined with MLP neural network is superior to other methods in the accuracy, recall, F1 score and AUC for ARDS mortality risk prediction.

4. Conclusions

This paper proposes an interpretability method for ARDS mortality risk prediction based on W-BDA and MLP. The extracted ARDS data was first subjected to data cleaning and k-nearest neighbor interpolation, and we provided an idea for dividing the source and target domains to prepare for transfer learning. Then XGBoost feature selection was adopted for the two domains to reduce dimensionality, eliminate redundant features, and combine SHAP to provide reliable explanations for medical staff. After that, W-BDA was used to map the feature space of the two domains, which avoided the disadvantage of not obtaining sufficient information due to insufficient samples and realized the weighted equilibrium adaptation of the conditional distribution and the marginal distribution. Finally, combined with the MLP network model constructed in this paper, it could deal with small sample data sets and achieve reliable ARDS mortality risk prediction. The method proposed in this paper has achieved significant improvements and has a certain degree of interpretability. In future work, the ARDS time-series metrics are sampled at intervals to expand the features of the two domains and further improve the generalization of our method.

R E F E R E N C E S

[1]. *S. Wang, L. Jiang, Y. Yang*, Transfer learning of medical image segmentation based on optimal transmission feature selection, Journal of Jilin University (Eng. Edition), 2022.

[2]. *M. W. Sjoding, D. Taylor, J. Motyka, et al*. "Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation", The Lancet Digital Health, **vol. 3**, no. 6, Jun. 2021, pp. e340-e348.

[3]. *L. Papazian, C. Aubron, L. Brochard, et al*, "Formal guidelines: management of acute respiratory distress syndrome", Annals of Intensive Care, **vol. 9**, no. 1, Jun. 2019, pp. 1-18.

[4]. *B. Huang, D. Liang, R. Zou, et al*, "Mortality prediction for patients with acute respiratory distress syndrome based on machine learning: a population-based study", Annals of Translational Medicine, **vol. 9**, no. 9, May. 2021, pp. 794-805.

[5]. *J. Xia, S. Pan, M. Yan, et al*, "A small-sample prognosis model for severe diseases based on transfer learning", Journal of Biomedical Engineering, **vol. 37**, no. 1, Jan. 2020, pp. 1-9.

[6]. *C. Jing, S. Sun, D. Qin*, "Establishment of an early risk prediction model for patients with acute respiratory distress syndrome", Chinese Journal of Nursing, **vol. 55**, no. 9, Sept. 2020, pp. 1285-1291.

[7]. *S. Aktar, A. Talukder, M. Ahamad, et al*, "Machine learning approaches to identify patient comorbidities and symptoms that increased risk of mortality in COVID-19", Diagnostics, **vol. 11**, no. 8, Jul. 2021, pp. 1383-1400.

[8]. *F. Xie, Q. Gao, C. Jin, et al*, "Hyperspectral image classification based on superpixel pooling convolutional neural network with transfer learning", Remote Sensing, **vol. 13**, no. 5, Mar. 2021, pp. 930-945.
[9]. *M. Pourhomayoun, M. Shakibi*, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making", Smart Health, **vol. 20**, Jan. 2021, pp. 100178-100185.
[10]. *M. Hu, X. Chen, Y. Sun, et al*, "Disease prediction model based on dynamic sampling and transfer learning", Journal of Computer Science, **vol. 42**, no. 10, Mar. 2019, pp. 2339-2354.
[11]. *H. Cao, Y. Zhang, B. Wu, et al*, "Prediction of complications of liver transplantation based on migration component analysis and support vector machine", Computer Applications, **vol. 41**, no. 12, Jul. 2021, pp. 3608-3613.
[12]. *J. Xu, C. Xu, B. Zou, et al*, "New incremental learning algorithm with support vector machines", IEEE Transactions on Systems, Man, and Cybernetics: Systems, **vol. 49**, no. 11, Nov. 2018, pp. 2230-2241.
[13]. *G. Liang, L. Zheng*, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis", Computer methods and programs in biomedicine, **vol. 187**, Apr. 2020, pp. 104964-104972.
[14]. *M. Long, J. Wang, G. Ding, et al*, "Transfer feature learning with joint distribution adaptation", Proceedings of the IEEE international conference on computer vision, 2013, pp. 2200-2207.
[15]. *S. A. Hicks, J. L. Isaksen, V. Thambawita, et al*, "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis", Scientific reports, **vol. 11**, no. 1, May. 2021, pp. 1-11.
[16]. *J. Wang, Y. Chen, S. Hao, et al*, "Balanced distribution adaptation for transfer learning", IEEE international conference on data mining (ICDM), Nov. 2017, pp. 1129-1134.
[17]. *Y. Yang, J. Guo, Q. Ye, et al*, "A weighted multi-feature transfer learning framework for intelligent medical decision making", Applied Soft Computing, **vol. 105**, Jul. 2021, pp. 107242-107252.
[18]. *Y. Shen, G. Cai, S. Chen, et al*, "Fluid intake-related association between urine output and mortality in acute respiratory distress syndrome", Respiratory research, **vol. 21**, no. 1, Jan. 2020, pp. 1-8.
[19]. *Y. Huang, X. Qin, Y. Chen, et al*, "Interpretability analysis of sepsis prediction model using LIME", Computer Applications, **vol. 41**, no. S1, Jun. 2021, pp. 332-335.
[20]. *L. Singhal, Y. Garg, P. Yang, et al*, "eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset Acute Respiratory Distress Syndrome (ARDS) among critically ill adults with COVID-19", PloS one, **vol. 16**, no. 9, Sept. 2021, pp. e0257056-e0257072.
[21]. *Y. Zhou, P. Chen, N. Liu, et al*, "Graph-Embedding Balanced Transfer Subspace Learning for Hyperspectral Cross-Scene Classification", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **vol. 15**, Mar. 2022, pp. 2944-2955.
[22]. *M. Desai, M. Shah*, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)", Clinical eHealth, **vol. 4**, Dec. 2021, pp. 1-11.
[23]. *Y. Shen, G. Cai, S. Chen, et al*, "Fluid intake-related association between urine output and mortality in acute respiratory distress syndrome", Respiratory research, **vol. 21**, no. 1, Jan. 2020, pp. 1-8.
[24]. *Y. Chi, H. He, Y. Long*, "A simple method of mechanical power calculation: using mean airway pressure to replace plateau pressure", Journal of Clinical Monitoring and Computing, **vol. 35**, no. 5, Oct. 2021, pp. 1139-1147.