Received: 2 April 2022 | Accepted: 16 August 2022

DOI: 10.1002/int.23055

RESEARCH ARTICLE

WILEY

A hybrid adaptive approach for instance transfer learning with dynamic and imbalanced data

Xiangzhou Zhang[1] | **Kang Liu**[1,2] | **Borong Yuan**[1,3] | **Hongnian Wang**[1,2] | **Shaoyong Chen**[1,3] | **Yunfei Xue**[1,3] | **Weiqi Chen**[1] | **Mei Liu**[4] | **Yong Hu**[1]

[1]Big Data Decision Institute, Jinan University, Guangzhou, China

[2]School of Management, Jinan University, Guangzhou, China

[3]College of Information Science and Technology, Jinan University, Guangzhou, China

[4]Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, Kansas, USA

Correspondence
Mei Liu, Division of Medical Informatics, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, USA.

Email: meiliu@kumc.edu
Yong Hu, Big Data Decision Institute, Jinan University, 601 Huangpu Road West, Guangzhou 510632, China.
Email: yonghu@jnu.edu.cn

Funding information
Major Research Plan of the National Natural Science Foundation of China, Grant/Award Number: 91746204; Science and Technology Planning Project of Guangdong Province, Grant/Award Number: 2017B030308008;

Abstract

Machine learning has demonstrated success in clinical risk prediction modeling with complex electronic health record (EHR) data. However, the evolving nature of clinical practices can dynamically change the underlying data distribution over time, leading to model performance drift. Adopting an outdated model is potentially risky and may result in unintentional losses. In this paper, we propose a novel *H*ybrid *A*daptive *B*oosting approach (HA-Boost) for transfer learning. HA-Boost is characterized by the domain similarity-based and class imbalance-based adaptation mechanisms, which simultaneously address two critical limitations of the classical TrAdaBoost algorithm. We validated HA-Boost in predicting hospital-acquired acute kidney injury using real-world longitudinal EHRs data. The experiment results demonstrate that HA-Boost stably outperforms the competing baselines in terms of both Area Under Receiver Operating Characteristic and Area Under Precision-Recall Curve across a 7-year time span. This study has confirmed the effectiveness of transfer learning

Xiangzhou Zhang, Kang Liu and Borong Yuan contributed equally to this study.

National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R01DK116986; National Science Foundation, Grant/Award Number: 2014554

as a superior model updating approach in a dynamic environment.

KEYWORDS
clinical risk prediction, electronic health records, imbalanced data, performance drift, transfer learning

1 | INTRODUCTION

With the worldwide adoption of electronic health record (EHR) systems, machine learning has made great strides in the secondary use of EHR data toward more accurate clinical risk prediction.[1–5] These studies have a fundamental assumption that the data distributions of the training and test sets are the same, and thus prediction model development is typically a one-time activity. However, clinical practices such as patient care and hospital conditions can change over time, and disease prevalence and cause can also change over time[6]; both cases would lead to changes in the data distributions, resulting in model performance drift.[7,8] Previously developed models can be outdated, making more and more inaccurate predictions for incoming patients.

Effective model updating is critical to continually maintaining prediction performance, and it requires a certain amount of new data (usually only a small amount of new data is available).[8,9] A range of model updating approaches is available, including recalibration, model-specific adaptation (e.g., incremental learning for tree-based model and neural network), model extension (e.g., incorporating new features), and full retraining, varying in analytical and computational complexity and sample size requirement.[10] The most straightforward method is full retraining periodically, whereas the corresponding challenge is how to prioritize the knowledge between new and old data. If the old data are completely ignored and the new (and often smaller) data are used alone, model overfitting would become a critical risk. Compared with full retraining, adapting the current model to a new environment requires a more complex learning paradigm, such as online learning[11] and lifelong learning,[12] both of which are prominent for dynamic environments or concept drift scenarios, while lifelong learning is superior to online learning in preserving old knowledge against being overwritten by new knowledge.[12] However, they are not suitable for EHR-based clinical prediction modeling because (1) they are designed to tackle streaming data (such as stock price, sensor, and other time-series data), that is, learn from a sequence of data instances one by one at each time and continually adapt the current model to the new environment, and (2) most EHR-based modeling studies are based on cross-sectional and longitudinal data of irregular time points.

Considering the above characteristics of EHR-based clinical prediction modeling, in the present study, we preferred another method named transfer learning, which is a prevalent machine learning approach toward modeling on a small target data set by (selectively) reusing source instances, features, and models/parameters.[13] We introduce transfer learning to address the problem of time-varying distribution, by treating the new data as the target domain and the historical data as the source domain, which can effectively reuse the historical data and only requires a small amount of new data. We believe that transfer learning can provide insights

from another perspective into the performance drift issue compared with other common approaches, such as recalibration and incremental learning. For example, when the data distribution significantly changes, transfer learning can immediately discard the old knowledge/model and reselect sample from the source domain for training, while incremental learning suffers from slow progressive adaptation.

Among various transfer learning algorithms, we chose the prevalent TrAdaBoost,[14] which is prominent on instance transfer, that is, reusing instances from the source domain to reduce the sample requirement of the target domain. Under the time-varying scenario, it iteratively assigns higher weights to the old data instances that are similar to the new data and lower weights to those not similar. During this process, old knowledge can be selectively reused when learning new knowledge. However, TrAdaBoost does not differentiate the causes of misclassifying the old data instances during training iterations,[15] such as (1) whether the misclassified instance is under a different distribution from the new data, (2) whether the weak classifier suffers from model underfitting and/or class imbalance issues. TrAdaBoost simply assumes all misclassified old data instances are under different distributions from the new data, and lowers their weights. To address these limitations, we propose a new set of reweighting criteria: when an old data instance is misclassified, we decrease its weight only if it is under a different distribution from the new data, while we increase its weight if it is under the same distribution as the new data. In addition, to reduce the impact of class imbalance on TrAdaBoost, we further propose a method to reweight the positive and negative instances differently during the training process. Because the instance reweighting mechanism of TrAdaBoost is very unfavorable to the minority class, it is prone to erroneously decrease the weight of misclassified minority class instances in old data, compared with the case of new data. In summary, we improved the original TrAdaBoost algorithm by introducing two adaptation mechanisms, named domain similarity-based adaptation and class imbalance-based adaptation, respectively.

We validated the proposed method for predicting hospital-acquired acute kidney injury (AKI). AKI is a highly lethal clinical syndrome caused by multiple etiologies in hospitalized patients. It is associated with much higher in-hospital morbidity and mortality, and the AKI survivors are at increased risk for developing chronic kidney disease, end-stage renal disease, and recurrent AKI.[8,16] Most importantly, AKI has a high nonrecognition rate.[17,18] Thus, early prediction of AKI has been an urgent demand. Early prediction of AKI can support clinical decision-making,[19,20] so that patient care can be assessed, intervened, and managed in advance.

To this end, we summarize the contributions of this study as follows:

1. We improve the classical transfer learning algorithm TrAdaBoost, by introducing the domain similarity-based and the class imbalance-based adaptation mechanisms. Essentially, these two mechanisms realize a superior instance reweighting procedure, addressing the issues of misclassification and class imbalance simultaneously.
2. We empirically validate the proposed approach on real-world EHR data, addressing the performance drift problem of the AKI prediction model over a long-time span. Particularly, we confirm the effectiveness of transfer learning as a superior model updating approach in a dynamic environment, such as clinical decision-making.

2 | METHOD

2.1 | Preliminary

We first briefly introduce the background of AdaBoost and TrAdaBoost, from which our proposed approach was derived.

AdaBoost, short for Adaptive Boosting, is a traditional machine learning method.[21] Its basic idea is paying more attention to instances that are hard to classify and less on others already handled well. Formally, let $\mathcal{X}$ and $\mathcal{Y}$ be the instance and label space, respectively. $\mathcal{S} \subseteq \{\mathcal{X} \times \mathcal{Y}\}$ be training instances. AdaBoost learning is a sequential process of training a set of base classifiers and then weighting them together. Assume the distribution of instance weights as w^t in tth round of learning. A base classifier as $h_t : \mathcal{X} \to \mathcal{Y}$ is generated from $\mathcal{S}$ and w^t. In the next round, w^{t+1} will be adjusted, where the weights of the misclassified instances by h_t will be increased. By using training instances $\mathcal{S}$ and the updated weights distribution w^{t+1}, a new base classifier h_{t+1} will be obtained. After repeating such a process for N rounds, the final discrimination function is derived by the weighted sum of all N base classifiers, where the weights of the classifiers are computed according to the error rate of the rounds.

TrAdaBoost is an extension of AdaBoost for transferring knowledge from the source domain to the target domain.[14] It assumes that the source and target domain data belong to the same feature and label space, but those data distributions are different. Due to differences in distribution, it considers there are some data in the source domain that can still be reused in learning for the target domain. To find them, TrAdaBoost tries to iteratively reweight the source domain data to put more weight on instances with the same distribution as the target domain and less on different. Formally, let $\mathcal{X}_s$ and $\mathcal{X}_d$ denote the same- and diff-distribution instance space, respectively. $\mathcal{Y}$ denotes the label space. The training instances are set to $\mathcal{S} \subseteq \{\mathcal{X} \times \mathcal{Y}\}$, where $\mathcal{X} = \mathcal{X}_d \cup \mathcal{X}_s$. Note that, $\mathcal{X}_d$ and $\mathcal{X}_s$ actually denote the source domain data and the target domain, respectively. In each learning round, TrAdaBoost decreases the weights of misclassified instances in the source domain, since its fundamental assumption is that this part of instances with diff-distribution from the target domain. The weight update is based on $Hedge(\beta)$,[21] decreasingly multiplied by $\beta^{|h_t(x)-c(x)|}$, where $c(x)$ is the mapping function from $\mathcal{X}$ to $\mathcal{Y}$. For the data of $\mathcal{X}_s$, the weights are updated in the same way as AdaBoost. Unlike AdaBoost, TrAdaBoost only unites the second half $N/2$ base classifiers to output the final decision.

2.2 | Prediction task and challenges

In this study, we consider AKI prediction as a binary classification task under a long-time span. We define the encounters that occurred AKI as the positive class and others as the negative class. Note that, the number of negative encounters greatly exceeded the number of positive encounters. Each encounter can be represented as $\mathbf{x} \in \mathbb{R}^d$ based on the observation records before the AKI prediction point, where d refers to the dimensionality of a set of variables of interest for a patient, usually with a large value. Our target population is new encounters. Thus, a large number of historical encounters and a small number of more recent encounters constitute the instance spaces $\mathcal{X}_d$ and $\mathcal{X}_s$, respectively. Note that, in continuous time, $\mathcal{X}_d$ and $\mathcal{X}_s$ change dynamically as new data are generated. The task is learning a model from the available data $\mathcal{X} = \mathcal{X}_d \cup \mathcal{X}_s$ to predict the class of a given unknown encounter representation $\mathbf{x}$.

Applying traditional machine learning approaches in such a scenario would suffer from the challenges of data shift and class imbalance, which would lead to performance drift.

2.3 | HA-Boost

To solve the above challenges, we propose the *H*ybrid *A*daptive *B*oosting approach (HA-Boost) as shown in Figure 1, where Figure 1A depicts the domain similarity-based and class imbalance-based adaptation mechanisms, for instance, weight updating, and Figure 1B represents the boosting learning process.

2.3.1 | Domain similarity-based adaptation

The TrAdaBoost approach intrinsically assumes that the source domain is under a significantly different distribution from the target domain, which is not suitable for the case of similar source and target domains, as well as the case of changing similarity. For this sake, we intuitively designed the domain similarity-based adaptation mechanism that balances the weight updating mechanisms of TrAdaBoost and AdaBoost in reweighting the instances of the source domain. The dominant mechanism depends on the distribution similarity between the source and target domains, that is, the higher the domain similarity, the greater the impact of the AdaBoost mechanism than the TrAdaBoost mechanism, and vice versa.

The first key point is how to measure the similarity between the source and target domains. Inspired by the idea of multiview learning, where one of the optimization objectives is to achieve

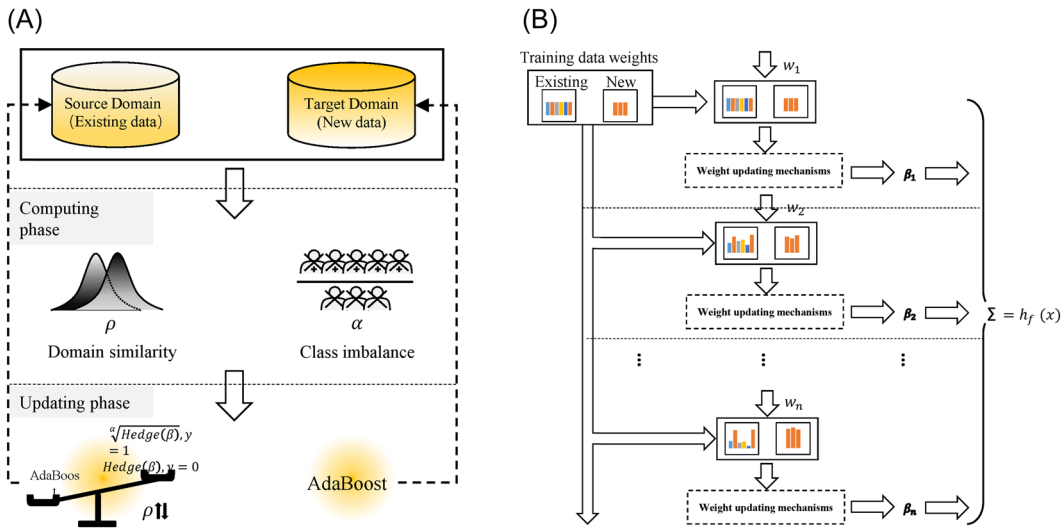

FIGURE 1 Framework of HA-Boost. (A) The weight updating mechanisms, where ρ is the distribution similarity measure, α is the balance factor for class imbalance-based adaptation, computed as the ratio of misclassified positive instances to misclassified negative instances, and β is used in $Hedge(\beta)$ for updating the weights. (B) The boosting learning process. HA-Boost, *H*ybrid *A*daptive *B*oosting approach. [Color figure can be viewed at wileyonlinelibrary.com]

consistent predictions for the same instance between classifiers from different views.[22] We calculated the similarity between the source and target domains by comparing the predictions of the source and target domain classifiers for the target domain instances. Formally, suppose classifier h^{d} and classifier h^{s} are trained based on $\mathcal{X}_{\mathrm{d}}$ and $\mathcal{X}_{\mathrm{s}}$, respectively. The round index is omitted for clarity. Note that, we performed 10-fold cross-validation on $\mathcal{X}_{\mathrm{s}}$ when constructing classifier h^{s}. Suppose $\mathcal{X}_{\mathrm{s}}$ is split into training set $\mathcal{X}_{\mathrm{s}}^{\mathrm{tra}} \in \mathbb{R}^{\frac{9m}{10} \times d}$ and validation set $\mathcal{X}_{\mathrm{s}}^{\mathrm{val}} \in \mathbb{R}^{\frac{m}{10} \times d}$, only $\mathcal{X}_{\mathrm{s}}^{\mathrm{tra}}$ is available for training. Besides, we applied stratified random sampling on $\mathcal{X}_{\mathrm{s}}^{\mathrm{tra}}$ to balance the proportion of positive and negative instances before training classifier h^{s}. The similarity ρ is defined as follows:

$$\rho = \frac{\sum_{x_i \in \mathcal{X}_{\mathrm{s}}^{\mathrm{val}}} w_i \left(h^{\mathrm{d}}(x_i) - \overline{h^{\mathrm{d}}\left(\mathcal{X}_{\mathrm{s}}^{\mathrm{val}}\right)}\right)\left(h^{\mathrm{s}}(x_i) - \overline{h^{\mathrm{s}}\left(\mathcal{X}_{\mathrm{s}}^{\mathrm{val}}\right)}\right)}{\sqrt{\sum_{x_i \in \mathcal{X}_{\mathrm{s}}^{\mathrm{val}}} w_i \left(h^{\mathrm{d}}(x_i) - \overline{h^{\mathrm{d}}\left(\mathcal{X}_{\mathrm{s}}^{\mathrm{val}}\right)}\right)^2 \sum_{x_i \in \mathcal{X}_{\mathrm{s}}^{\mathrm{val}}} w_i \left(h^{\mathrm{s}}(x_i) - \overline{h^{\mathrm{s}}\left(\mathcal{X}_{\mathrm{s}}^{\mathrm{val}}\right)}\right)^2}}, \tag{1}$$

where w_i is an instance weight, $\overline{h^{\mathrm{d}}(\mathcal{X}_{\mathrm{s}}^{\mathrm{val}})} = \frac{m}{10}\sum_{x_i \in \mathcal{X}_{\mathrm{s}}^{\mathrm{val}}} w_i h^{\mathrm{d}}(x_i)$, and $\overline{h^{\mathrm{s}}(\mathcal{X}_{\mathrm{s}}^{\mathrm{val}})} = \frac{m}{10}\sum_{x_i \in \mathcal{X}_{\mathrm{s}}^{\mathrm{val}}} w_i h^{\mathrm{s}}(x_i)$. In practice, we use the mean value produced by the 10-fold cross-validation to reduce variance.

The next key point is how to assign the impact of TrAdaBoost and AdaBoost on weight updating based on similarity ρ. We propose the following formula to update the weight of instance in $\mathcal{X}_{\mathrm{d}}$,

$$w_i^{t+1} = \begin{cases} \underbrace{w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|}}_{\sim\text{AdaBoost}}, & f(\rho^t) \geq p, \\ \underbrace{2(1-\varepsilon_t) w_i^t \beta^{|h_t(x_i)-c(x_i)|}}_{\sim\text{TrAdaBoost}}, & f(\rho^t) < q, \\ \underbrace{f(\rho^t) * w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|}}_{\sim\text{AdaBoost}} \\ + \underbrace{(1-f(\rho^t)) * 2(1-\varepsilon_t) w_i^t \beta^{|h_t(x_i)-c(x_i)|}}_{\sim\text{TrAdaBoost}}, & \text{else}, \end{cases} \tag{2}$$

where $\beta = 1/(1+\sqrt{2\ln n/N})$, $\beta_t = \epsilon_t/(1-\epsilon_t)$, and ε_t is the error of h_t on $\mathcal{X}_{\mathrm{s}}$, calculated by

$$\varepsilon_t = \sum_{i=n+1}^{n+m} \frac{w_i^t \cdot |h_t(x_i) - c(x_i)|}{\sum_{i=n+1}^{n+m} w_i^t}. \tag{3}$$

In Equation (2), $f(\cdot)$ is a mapping function defined as $f(x) = \frac{x-q}{p-q}$, where p and q are the two hyperparameters to normalize the similarity. When the similarity is greater than p, we consider that the source and target domains belong to the same distribution, and less than q, different distributions. For brevity, we redefine f as follows:

$$f(x) = \begin{cases} 1, & x \geq p, \\ \frac{x-q}{p-1}, & q < x < p, \\ 0, & x \leq q, \end{cases} \tag{4}$$

then Equation (2) is modified as follows:

$$w_i^{t+1} = f(\rho^t) * w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|} + (1 - f(\rho^t)) * 2(1-\varepsilon_t) w_i^t \beta^{|h_t(x_i)-c(x_i)|}. \tag{5}$$

For the instance from $\mathcal{X}_s$, AdaBoost is still good and so is used to update

$$w_i^{t+1} = w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|}. \tag{6}$$

The ensemble classifier is similar to TrAdaBoost, based on the voting of the base classifiers obtained from the training process. TrAdaBoost considers that the distributions between the source and target domains are more different from the beginning, so only the second half of the base classifiers are considered for voting. However, we consider that the similarity between the source and target domains is uncertain, so we use all base classifiers.

2.3.2 | Class imbalance-based adaptation

Class imbalance is a very common problem in machine learning. Many models do not handle this problem very well, since their training process usually prefers the majority class over the minority class. However, the minority class usually carries richer concepts than the majority class, especially for medical data.[23] TrAdaBoost is more extreme, where it decreases the weight of diff-distribution instances if they are misclassified. Even though the minority class instances are more important, they are easily misclassified in an imbalanced data scenario, thus easily ignored by TrAdaBoost. To solve this drawback, we designed the class imbalance-based adaptation mechanism to promote TrAdaBoost in learning diff-distribution instances.

We used α_t as a balance factor for class imbalance-based adaptation. In detail, suppose ϵ^- and ϵ^+ represent the error measures of negative and positive classes in the source domain under h_t prediction, respectively, and n^- and n^+ represent the number of instances they have, respectively. On the basis of Equation (3), the balance factor α_t is defined as follows:

$$\alpha_t = \frac{n^+}{n^-}\frac{\epsilon_t^-}{\epsilon_t^+} = \frac{n^+}{n^-}\frac{\sum_{n^-} \frac{w_i^t \cdot |h_t(x_i) - c(x_i)|}{\sum_{n^-} w_i^t}}{\sum_{n^+} \frac{w_i^t \cdot |h_t(x_i) - c(x_i)|}{\sum_{n^+} w_i^t}}. \tag{7}$$

On the basis of the weight update scheme of TrAdaBoost and the balance factor α, we adjusted the reweighting scheme of positive instances. If a positive instance x^+ from diff-distribution is wrongly classified, its weight is multiplied by $\beta^{\frac{|h_t(x)-c(x)|}{\alpha_t}}$ rather than $\beta^{|h_t(x)-c(x)|}$. As a result, Equation (5) is modified as follows:

$$w_i^{t+1} = f(\rho^t) * w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|} + (1 - f(\rho^t)) * 2(1-\varepsilon_t) w_i^t \beta^{\lambda|h_t(x_i)-c(x_i)|},$$

$$\text{where} \quad \lambda = \begin{cases} 1, & c(x_i) = 0, \\ \alpha^{-1}, & c(x_i) = 1. \end{cases} \tag{8}$$

Algorithm 1 outlines the details of our novel modeling approach, including two adaptation mechanisms. First, it measures the similarity of $\mathcal{X}_d$ and $\mathcal{X}_s$. Then it concatenates $\mathcal{X}_d$ and $\mathcal{X}_s$ to train a classifier h_t. Finally, the instance weights are adjusted according to the similarity of $\mathcal{X}_d$

and $\mathcal{X}_s$ and the error rate of h_t. After N rounds, it integrates all base classifiers into an ensemble classifier.

Algorithm 1. *HA-Boost Algorithm*

Input: $\mathcal{X}_d \in \mathbb{R}^{n\times d}$ and $\mathcal{X}_s \in \mathbb{R}^{m\times d}$, base classifier algorithm, and the maximum number of iterations N;

Output: The final classifier $h_f(x)$;
Initialize: The initial weight vector, that
$w^1 = (w_1^1, ..., w_{(n+m)}^1)$, where

$$w_i^1 = \begin{cases} \frac{1}{n}, & 1 \le i \le n, \\ \frac{1}{m}, & n+1 \le i \le n+m; \end{cases}$$

for $t = 1, ..., N$ **do**
Normalize weights $w^t = w^t/(\sum_{i=1}^{n+m} w_i^t)$;
Calculate the similarity ρ_t between $\mathcal{X}_d$ and $\mathcal{X}_s$ by Equation (1);
Call the base classifier algorithm over $\mathcal{X}_d$ and $\mathcal{X}_s$, and then get back a base classifier $h_t(x)$;
Calculate the error of base classifier ε_t on $\mathcal{X}_s$ by Equation (3);
Calculate the balance factor α_t on $\mathcal{X}_d$ by Equation (7);
if $n+1 \le i \le n+m$ **then**
Update w_i^t by Equation (6);
else
Update w_i^t by Equation (8);
end
end
Output: an ensemble classifier,

$$h_f(x) = \begin{cases} 1, & \prod_{t=1}^{N} \beta_t^{-h_t(x)} \ge \prod_{t=1}^{N} \beta_t^{-\frac{1}{2}}, \\ 0, & \textbf{otherwise} \end{cases}$$

3 | EXPERIMENT

3.1 | Study population

To evaluate HA-Boost, we conducted experiments on a large-scale retrospective observational cohort constructed in our previous study on AKI prediction,[5] drawn from a deidentified EHR repository at the University of Kansas Medical Center (KUMC). The data set contains adult patients (age ≥ 18 years at admission) who were hospitalized for at least 2 days from 2010 to 2017. We defined AKI using the Kidney Disease Improving Global Outcomes (KDIGO) serum creatinine (SCr) criteria[24]: (1) an increase in SCr of 0.3 mg/dl (26.5 mol/L) within 48 h, or (2) an increase in SCr of 1.5 times of the patient's baseline creatinine (which is defined as the most recent SCr, or admission SCr value when past measurements were not available). According to the necessary conditions for determining AKI, we excluded patients that (1) with <2 SCr measurement, or (2) with moderate-to-severe kidney dysfunction at the time of admission, that is, estimated glomerular filtration rate lower than 60 ml/min/1.73 m^2 using the Modified Died in Renal Disease equation, or SCr > 1.3 mg/dl within 24 h of admission. The final cohort consisted of 141,696 encounters.

3.2 | Data preprocessing

The original data cover more than 28,000 variables. Each instance has a sequence of time-stamped clinical variables, including (1) demographics, such as age, race, and gender; (2) vital signs, such as BMI, tobacco, and blood pressure; (3) laboratory tests based on LOINC codes; (4) comorbidities based on ICD-9 and ICD-10 codes; and (5) medications based on RXNORM and NDC codes.[25] When constructing the training instance, each encounter representation is generated from the clinical variables collected before the prediction point, and its label is whether the sample outbreaks AKI within the next h hours. In this study, we set h to 24 and 48 h for predicting AKI 24 and 48 h ahead, respectively. Table 1 shows the patient demographic characteristics in different years for 24 h prediction task. As can be seen that the number of negative instances is almost six times the positive instances in all years, which indicates the data set has an imbalance problem. Also, the distribution of each feature changes across years.

We represent each instance as a vector of fixed dimensions, using the following criteria: (1) we selected the most recent value for the variables which are repeatedly measured within a certain time interval; (2) we applied one-hot encoding for categorical variables; (3) we set the number of times the patient took the medication before the prediction point as the medication representation; (4) we removed variables with a missing rate >99.99%; (5) we made feature selection for categorical and continuous variables by chi-square statistics and analysis of variance, respectively, and retain variables with $p < 0.05$; (6) we constructed additional variables, such as daily blood pressure trends, which are useful for AKI prediction.[26] Note that the variables directly related to SCr and blood urea nitrogen are not included, because they are used to determine the AKI outcome.[5]

3.3 | Experimental settings

The evaluation metrics used in this study include: (1) the Area Under Receiver Operating Characteristic (AUROC), which is the most widely used performance metric of binary classification tasks, to summarize the sensitivity and specificity of the model; (2) the Area Under Precision-Recall Curve (AUPRC), which is one of the most appropriate performance metrics of imbalance classification tasks, to summarize the sensitivity and precision of the model by focusing on how well the model performs on the positive class.

Considering that the available data size of the target domain is usually small in real-world applications due to scarcity of data, high costs of data cleaning, and so forth, we used all historical data as source domain data to build the new model. For example, to build a prediction model for the target domain 2014, we used all data up to 2013 as the source domain. Using stratified random sampling, we extracted 20% of the target domain data as training data to help capture the data shift in 2014, and the rest as test data. Fivefold cross-validation was adopted in all experiments to reduce variance.

To illustrate the effectiveness of our method, we chose AdaBoost which did not use historical data as the baseline to compare the model performances between transferring or not, and TrAdaBoost as another baseline to measure whether the model performance improved or not. For all models, we choose LightGBM as the base learner, which is widely utilized in various fields of machine learning-based researches, and we used the default parameters provided in the official release code (available at github.com/Microsoft/LightGBM) if not specified otherwise. We set the number of iteration parameter N to 5, hyperparameters p and q to 0.9 and 0.1, respectively.

TABLE 1 Patient demographic characteristics for 24 h AKI prediction task

Characteristics	2010	2011	2012	2013	2014	2015	2016	2017
Age								
18–25	869	886	923	918	1077	1082	1086	1001
	(5.81%)	(5.75%)	(5.53%)	(5.26%)	(5.76%)	(5.38%)	(5.32%)	(5.56%)
26–35	1290	1275	1468	1567	1717	1814	1823	1664
	(8.63%)	(8.27%)	(8.80%)	(8.98%)	(9.18%)	(9.03%)	(8.94%)	(9.24%)
36–45	1640	1727	1696	1861	1819	2136	2196	1919
	(10.97%)	(11.20%)	(10.17%)	(10.66%)	(9.73%)	(10.63%)	(10.77%)	(10.66%)
46–55	3025	2998	3203	3133	3150	3482	3259	2762
	(20.24%)	(19.44%)	(19.20%)	(17.95%)	(16.84%)	(17.33%)	(15.98%)	(15.34%)
56–65	3383	3659	3951	4161	4558	4897	4840	4088
	(22.63%)	(23.73%)	(23.68%)	(23.85%)	(24.37%)	(24.37%)	(23.73%)	(22.71%)
>65	4739	4877	5441	5810	6380	6683	7195	6568
	(31.71%)	(31.62%)	(32.62%)	(33.30%)	(34.12%)	(33.26%)	(35.27%)	(36.48%)
Sex								
Female	7399	7787	8250	8810	9394	9980	10,149	8957
	(49.50%)	(50.49%)	(49.45%)	(50.49%)	(50.23%)	(49.67%)	(49.75%)	(49.76%)
Male	7547	7635	8432	8639	9307	10,114	10,249	9045
	(50.50%)	(49.51%)	(50.55%)	(49.51%)	(49.77%)	(50.33%)	(50.24%)	(50.24%)
Race								
White	10,891	11,476	12,667	13,273	14,230	15,270	15,388	13,514
	(72.87%)	(74.41%)	(75.93%)	(76.06%)	(76.09%)	(75.99%)	(75.44%)	(75.07%)
Black	2286	2240	2255	2510	2685	2883	2896	2614
	(15.30%)	(14.52%)	(13.52%)	(14.38%)	(14.36%)	(14.35%)	(14.20%)	(14.52%)
Asian	125	128	153	167	210	184	254	149
	(0.84%)	(0.83%)	(0.92%)	(0.96%)	(1.12%)	(0.92%)	(1.25%)	(0.83%)
Native American	53	52	46	79	68	87	80	63
	(0.35%)	(0.34%)	(0.28%)	(0.45%)	(0.36%)	(0.43%)	(0.39%)	(0.35%)
Other	1591	1526	1561	1421	1508	1670	1781	1662
	(10.64%)	(9.90%)	(9.36%)	(8.15%)	(8.06%)	(8.31%)	(8.73%)	(9.24%)
Hispanic								
Yes	792	798	964	970	1064	1169	1155	1039
	(5.30%)	(5.17%)	(5.78%)	(5.56%)	(5.69%)	(5.82%)	(5.66%)	(5.77%)
No	14,139	14,572	15,655	16,383	17,473	18,716	19,028	16,766
	(94.60%)	(94.49%)	(93.84%)	(93.89%)	(93.43%)	(93.14%)	(93.28%)	(93.13%)

TABLE 1 (Continued)

Characteristics	2010	2011	2012	2013	2014	2015	2016	2017
Unkonwn	15 (0.10%)	52 (0.34%)	63 (0.38%)	97 (0.56%)	164 (0.88%)	209 (1.04%)	216 (1.06%)	197 (1.09%)
AKI								
Non-AKI	12,414 (83.06%)	12,937 (83.89%)	14,097 (84.50%)	15,124 (86.67%)	16,165 (86.44%)	17,435 (86.77%)	17,660 (86.57%)	15,705 (87.24%)
Any AKI	2532 (16.94%)	2485 (16.11%)	2585 (15.50%)	2326 (13.33%)	2536 (13.56%)	2659 (13.23%)	2739 (13.43%)	2297 (12.76%)

3.4 | Results

Figure 2 shows the model performances for 24 and 48 h AKI prediction tasks in terms of AUROC and AUPRC. First, predicting AKI in 24 h showed uniformly better performance than 48 h prediction, in terms of both AUROC and AUPRC. Second, the HA-Boost model performed best, with an average AUROC of 0.854 and 0.782 for predicting AKI risk in 24 and 48 h, respectively, and an average AUPRC of 0.615 and 0.440. Third, the HA-Boost model outperforms the TrAdaBoost model by an average ΔAUROC of 0.028 and 0.023/year for 24 and 48 h prediction tasks, respectively, and an average ΔAUPRC of 0.069 and 0.058. Four, the AdaBoost model, which only used training data from the target domain in our experiment setting, underperformed the TrAdaBoost and HA-Boost models.

To examine the effects of the proposed adaptation mechanisms, we conducted an ablation study by removing the components one by one, with the results shown in Figure 3. When removing the class imbalance-based adaptation mechanism from HA-Boost, that is, compared with HA-Boost, the TrAdaBoost with domain similarity-based adaptation mechanism suffered an average drop of AUROC by 0.003 and 0.006/year for predicting AKI risk in 24 and 48 h, respectively, and AUPRC by 0.006 and 0.009. When removing the domain similarity-based adaptation mechanism, that is, compared with HA-Boost, the TrAdaBoost with class imbalance-based adaptation mechanism suffered an average drop of AUROC by 0.002 and 0.004/year for predicting AKI risk in 24 and 48 h, respectively, and AUPRC by 0.033 and 0.025. Either the domain similarity-based or class imbalance-based adaptation mechanism could result in a considerable gain in AUROC and AUPRC compared with TrAdaBoost. Except for the AUROC of 2013, the combination of these two adaptation mechanisms always had a complementary effect, yielding better performance.

We also conducted experiments with different sizes of available training data from the target domain, that is, splitting the target domain data into training and testing data sets as 1:9, 1:4, 1:2, 1:1, and 2:1. Figure 4 shows the performance of the HA-Boost model, averaged by the cross-validation approach. As can be seen, whatever the size of training data from the target domain, predicting AKI in 48 h showed uniformly worse performance than the 24 h prediction. As the size of available training data from the target domain increased, the model performed better and better, which implied the general scarcity of training data (i.e., new data in the real-world scenario) from the target domain, and inherently the difference between source and target domain.

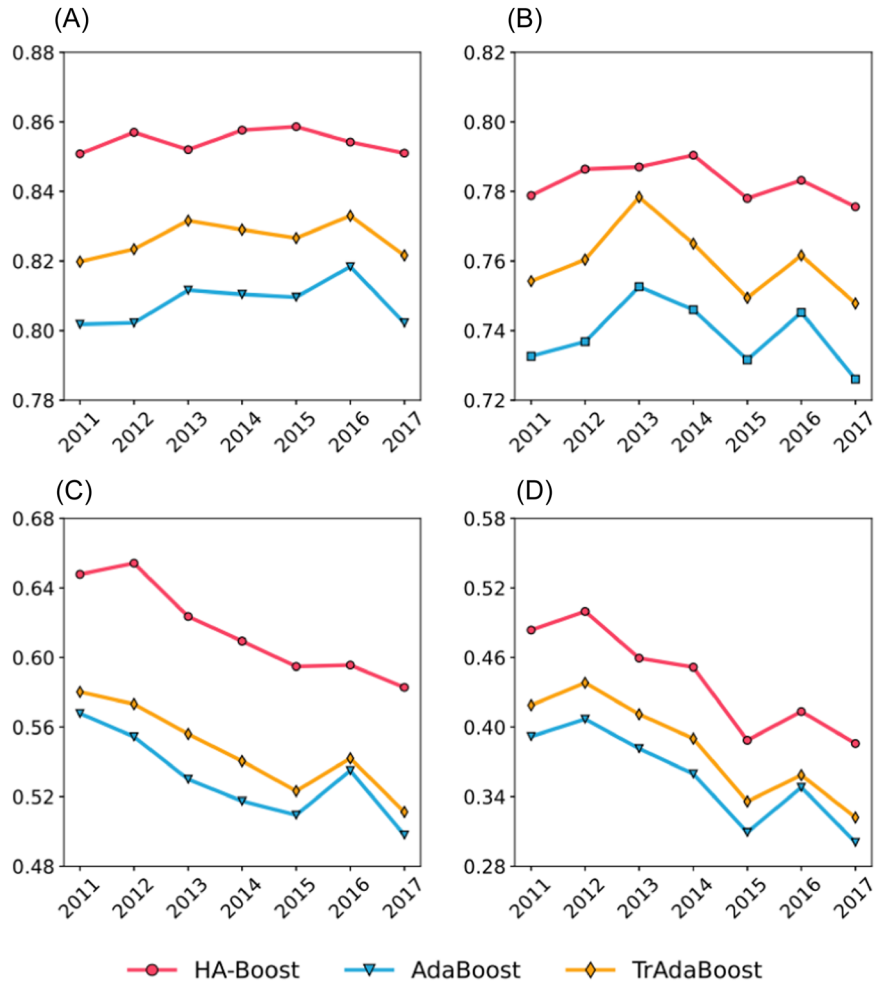

FIGURE 2 Model performance for AKI prediction. The *X*-axis is years. The *Y*-axis of both (A) and (B) is AUROC, and *Y*-axis of both (C) and (D) is AUPRC. AKI, acute kidney injury; AUPRC, Area Under Precision-Recall Curve; AUROC, Area Under Receiver Operating Characteristic; HA-Boost, Hybrid Adaptive Boosting approach. [Color figure can be viewed at wileyonlinelibrary.com]

To further investigate the impact of the class imbalance issue, we analyzed the classification error rates of positive and negative samples at each iteration during model training. Take the target domain 2014 as an example, Figure 5 shows the error rates of the original TrAdaBoost and the improved TrAdaBoost with the class imbalance-based adaptation mechanism. Beginning with the same error rate, the positive error rate of the improved TrAdaBoost with class imbalance-based adaptation gradually decreased, while the original TrAdaBoost gradually increased. On the contrary, the negative error rate had an opposite trend to the positive error rate. Besides, the improved TrAdaBoost with the class imbalance-based adaptation mechanism had a smaller magnitude of changes in error rates, compared with the original TrAdaBoost. Thus, the choice under this trade-off between positive and negative error rates is clear, since the positive samples receive more attention in clinical scenarios. Besides, we found that the performance differences between the 24 and 48 h prediction were mainly reflected in the positive error rate, whereas less difference reflected in the negative error rate.

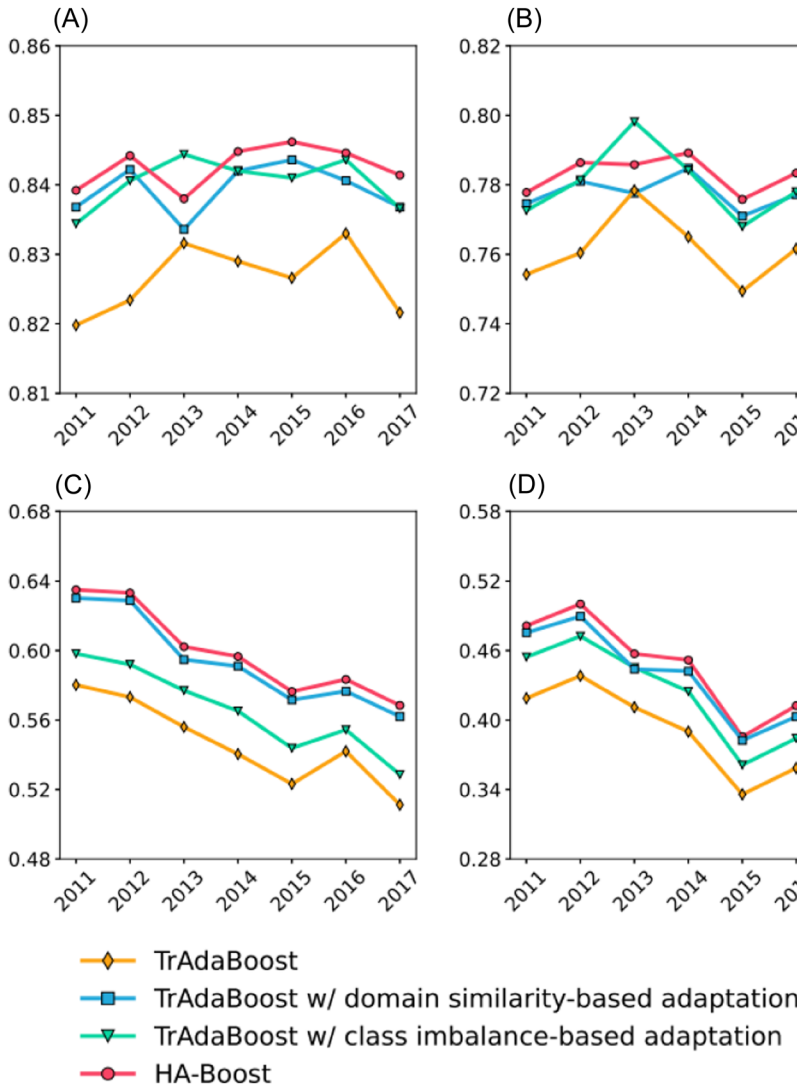

FIGURE 3 Ablation study of the impact of adaptation mechanisms on the AKI prediction model. The *X*-axis is years. The *Y*-axis of (A) and (B) is AUROC, and the *Y*-axis of (C) and (D) is AUPRC. AKI, acute kidney injury; AUPRC, Area Under Precision-Recall Curve; AUROC, Area Under Receiver Operating Characteristic; HA-Boost, Hybrid Adaptive Boosting approach. [Color figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

Under the scenario of data shift over time, it is theoretically necessary to treat the historical and new data in different ways. Simply pooling the historical and new data as training data to construct a prediction model is contradicted with this assumption. Thus, in this study, we trained the AdaBoost model only based on the new data, completely ignoring the historical data, to obtain a baseline model that only focuses on the current situation. Unsurprisingly, the AdaBoost model was vulnerable to the problem of insufficient new data for training, and yielded the worst performance in terms of AUROC and AUPRC. Nevertheless, we trained the

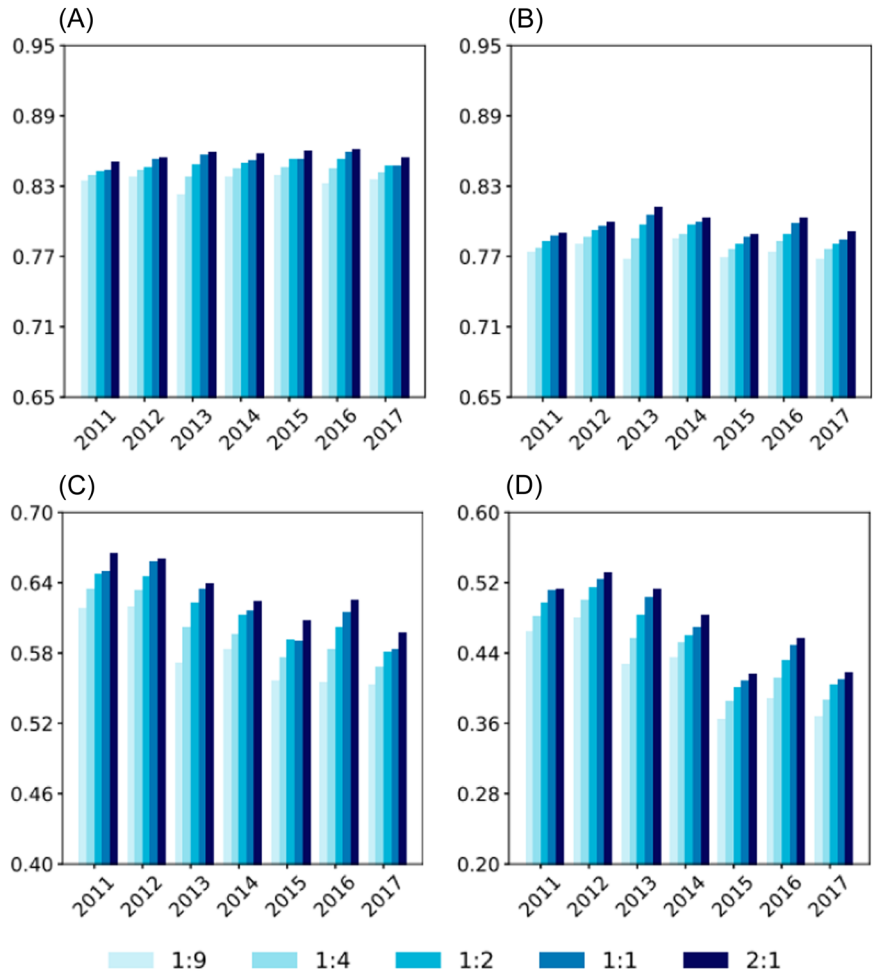

FIGURE 4 Comparison of HA-Boost model performances on different sizes of training data from target domain for 24 h and 48 h AKI prediction. The *X*-axis is years. The *Y*-axis of (A) and (B) is AUROC. The *Y*-axis of (C) and (D) is AUPRC. AKI, acute kidney injury; AUPRC, Area Under Precision-Recall Curve; AUROC, Area Under Receiver Operating Characteristic; HA-Boost, Hybrid Adaptive Boosting approach. [Color figure can be viewed at wileyonlinelibrary.com]

TrAdaBoost and HA-Boost models based on both the new and historical data, by iteratively overweighting the instances of historical data with the same distribution as the new data while underweighting the ones with different distributions. They succeeded in selectively reusing the historical data to increase the training sample size without being negatively affected by the outdated knowledge underlying the historical data, resulting in a significant performance gain. Compared with the TrAdaBoost model, the HA-Boost model has a superior ability of instance screening and thus uniformly performed better.

As for the HA-Boost model, both the domain similarity-based and class imbalance-based adaptation mechanisms were applied in each iteration during model training process. This intrinsically provided more dynamic characteristics to HA-Boost than the one-shot scheme of domain similarity measurement and class imbalance-orient resampling before model training. Furthermore, the ablation study showed that the domain similarity-based adaptation could

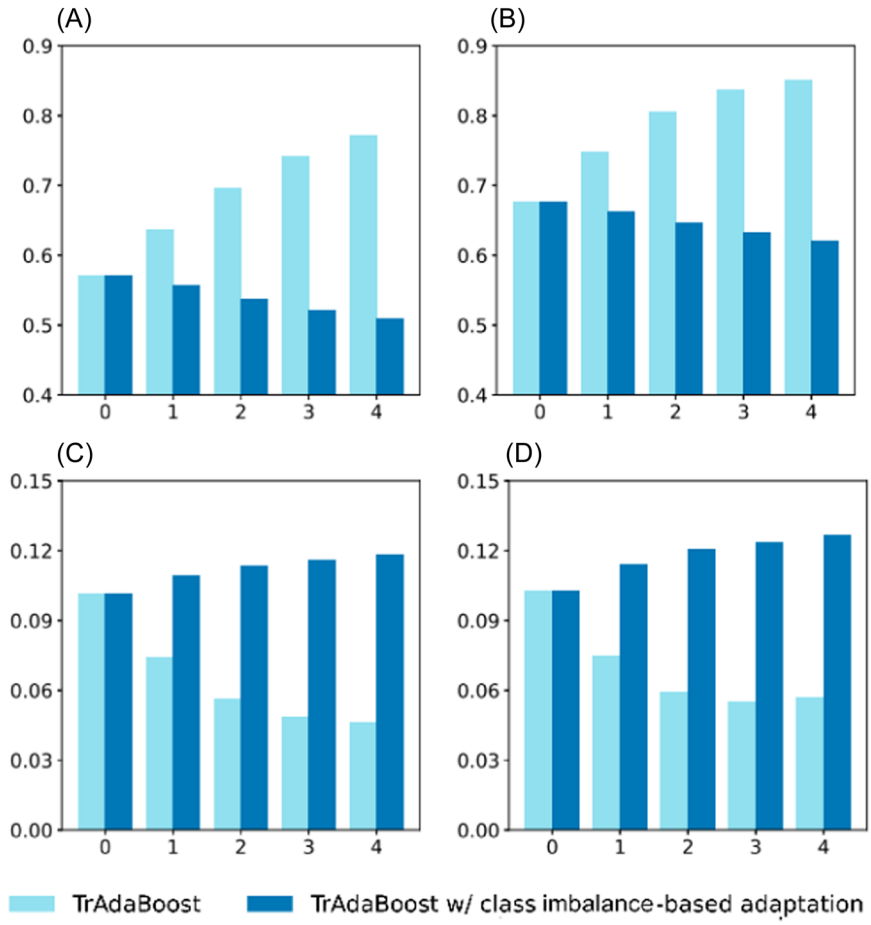

FIGURE 5 Positive and negative error rates for AKI prediction (Take 2014 as a case study). The *X*-axis is the iteration number. The *Y*-axis of (A) and (B) is a positive error rate, and the *Y*-axis of (C) and (D) is a negative error rate. AKI, acute kidney injury. [Color figure can be viewed at wileyonlinelibrary.com]

obtain more performance gain than the class imbalance-based adaptation mechanism (especially in terms of AUPRC), and achieved a very close performance to the HA-Boost model that incorporated both two adaptation mechanisms. This might imply that the domain difference (i.e., data shift between the historical and new data) was highly related to the changing class imbalance over time, and the domain similarity-based adaptation mechanism could partially capture these changes, having a similar effect as the class imbalance-based adaptation mechanism. Therefore, the proposed domain similarity-based adaptation mechanism would be the most important improvement to TrAdaBoost (at least for our study population).

Usually, the change in clinical practices would not be drastic and sharp, and thus the data shift over time would be in a gradual way. For example, as shown in Table 1, the difference in AKI incidence rate between any two consecutive years is relatively small, while the AKI incidence rate has a significant drop from 16.94% to 12.76% between 2010 and 2017. However, the TrAdaBoost approach intrinsically assumes that the distributions of target and source domains are significantly different, and adopts opposite instance reweighting methods between

the source and target domains, considering the misclassified instances in the source domain belong to a different distribution from the target domain and thus underweighted. In other words, in the scenario of gradual data shift, the TrAdaBoost approach might easily underweight or discard some misclassified instances in the source domain that actually belong to the same distribution as the target domain. And worse yet, the class imbalance problem would lead to the minority class being more likely to be misclassified. Unfortunately, the positive class of interest in most clinical modeling studies, for example, the AKI patients, belongs to the minority class. Therefore, the TrAdaBoost approach would not be expected to perform well in this case.

Next, we analyze the training process of the HA-Boost model. In the initial iterations, the source and target domains are similar and the AdaBoost-like reweighting scheme would dominate, thus the weights of misclassified instances in the source domain would be increased, which promotes them to be correctly classified in the subsequent iterations, while the weights of correctly classified instances in the source domain would be decreased. However, the misclassified instances in the source domain with a different distribution from the target domain also "erroneously" gain increased weights. As this continues, the source and target domains become more and more different, and the AdaBoost-like reweighting scheme would gradually lead to negative transfer.[27] Take 2014 for illustration, we found that the average similarity ρ dropped from 0.67 to 0.40 through five iterations. Therefore, in the later iterations, the TrAdaBoost-like reweighting scheme is more prone to dominate, thus the weights of misclassified instances in the source domain would be decreased while the weights of correctly classified instances would be increased. In summary, the HA-Boost approach can maintain a balance between the AdaBoost-like and TrAdaBoost-like reweighting schemes during the entire iterations.

As for the performance drift issue of AKI prediction models, Davis et al. have conducted the most comprehensive investigations.[8,10] However, their data set was collected from the US Department of Veterans Affairs, making it not a typical scenario of population drift, as reflected in the experiment results that model discrimination was maintained for all models but calibration declined.[8] They also found that machine learning models (random forest and neural network models) maintained more stable calibration compared with regression models. Furthermore, they focused on investigating a comprehensive procedure to select the appropriate updating method among several alternatives to correct performance drift by balancing simplicity against accuracy.[10] To the best of our knowledge, our HA-Boost model was the first to demonstrate the effectiveness of transfer learning in addressing the performance drift issue, especially for the AKI prediction task.

5 | CONCLUSION AND FUTURE WORK

Clinical risk prediction is significant for improving patient care, but prediction models are facing the problem of performance drift across years. Periodical model updating can address this problem, and it is necessary to consider how to balance the impact of the old and new data. In this paper, we designed two adaptation mechanisms, including domain similarity-based adaptation and class imbalance-based adaptation, to improve the classical TrAdaBoost. Using real-world EHRs data for AKI prediction model development, we have demonstrated that HA-Boost is more suitable for long-time span data-shift scenarios compared with TrAdaBoost. Furthermore, we have confirmed the effectiveness of transfer learning as a superior model

updating approach in a dynamic environment. However, we have only validated this on a single-center data set, without external validation in other medical centers, so that the generalizability of HA-Boost is still uncertain. Also, we still need to investigate the impact of AKI prediction models on clinical decision support to improve medical care in the future.

AUTHOR CONTRIBUTIONS

Yong Hu and Mei Liu initiated the project and designed the overall study. Mei Liu extracted the data used in this study. Xiangzhou Zhang and Kang Liu designed the algorithm. Shaoyong Chen, Yunfei Xue, and Borong Yuan designed the initial training and testing setup, and performed the first-round experiments. Hongnian Wang performed replication experiments and the second-round experiments, and drafted the paper, with critical revisions by Yong Hu, Mei Liu, Xiangzhou Zhang, Kang Liu, and Weiqi Chen.

ACKNOWLEDGMENTS

This study was supported by the Major Research Plan of the National Natural Science Foundation of China (Key Program, 91746204), the Science and Technology Development in Guangdong Province (Major Projects of Advanced and Key Techniques Innovation, 2017B030308008), and Guangdong Engineering Technology Research Center for Big Data Precision Healthcare (603141789047). Mei Liu was supported by NIH/NIDDK under award number R01DK116986 and the NSF under award number 2014554. The clinical data set used for analysis described in this study was obtained from the University of Kansas Medical Center (KUMC) HERON clinical data repository which is supported by institutional funding and by the KUMC Clinical Translational Science Award (CTSA) grant UL1TR002366 from NIH.

DATA AVAILABILITY STATEMENT

No. Research data are not shared.

ORCID

Xiangzhou Zhang https://orcid.org/0000-0003-3752-0045

REFERENCES

1. Du J, Zeng D, Li Z, et al. An interpretable outcome prediction model based on electronic health records and hierarchical attention. *Int J Intell Syst*. 2021;37:3460-3479. doi:10.1002/int.22697
2. Liao S, Jin L, Dai WQ, et al. A machine learning-based risk scoring system for infertility considering different age groups. *Int J Intell Syst*. 2021;36(3):1331-1344. doi:10.1002/int.22344
3. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-119. doi:10.1038/s41586-019-1390-1
4. Song X, Waitman LR, Yu AS, Robbins DC, Hu Y, Liu M. Longitudinal risk prediction of chronic kidney disease in diabetic patients using temporal-enhanced gradient boosting machine: retrospective cohort study. *JMIR Med Inf*. 2020;8(1):e15510. doi:10.2196/15510
5. Song X, Yu ASL, Kellum JA, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun*. 2020;11(1):5668. doi:10.1038/s41467-020-19551-w
6. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626
7. Martin GP, Sperrin M, Sotgiu G. Performance of prediction models for COVID-19: the Caudine Forks of the external validation. *Eur Respir J*. 2020;56(6):2003728. doi:10.1183/13993003.03728-2020
8. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inf Assoc*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

9. Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res*. 2021;5(1):1. doi:10.1186/s41512-020-00090-3
10. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inf Assoc*. 2019;26(12):1448-1457. doi:10.1093/jamia/ocz127
11. Hoi SCH, Sahoo D, Lu J, Zhao P. Online learning: a comprehensive survey. *Neurocomputing*. 2021;459:249-289. doi:10.1016/j.neucom.2021.04.112
12. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. *Neural Net*. 2019;113:54-71. doi:10.1016/j.neunet.2019.01.012
13. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191
14. Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Ghahramani Z, ed. *Proceedings of the 24th International Conference on Machine Learning—ICML '07*. ACM Press; 2007:193-200. doi:10.1145/1273496.1273521
15. Zheng L, Liu G, Yan C, Jiang C, Zhou M, Li M. Improved TrAdaBoost and its application to transaction fraud detection. *IEEE Trans Comput Soc Syst*. 2020;7(5):1304-1316. doi:10.1109/TCSS.2020.3017013
16. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA*. 2018;320(1):27-28. doi:10.1001/jama.2018.5602
17. Murugan R, Kellum JA. Acute kidney injury: what's the prognosis. *Nat Rev Nephrol*. 2011;7(4):209-217. doi:10.1038/nrneph.2011.13
18. Chawla LS, Bellomo R, Bihorac A, et al. Acute kidney disease and renal recovery: consensus report of the acute disease quality initiative (ADQI) 16 workgroup. *Nat Rev Nephrol*. 2017;13(4):241-257. doi:10.1038/nrneph.2017.2
19. Yang L, Xing G, Wang L, et al. Acute kidney injury in China: a cross-sectional survey. *Lancet*. 2015;386(10002):1465-1471. doi:10.1016/S0140-6736(15)00344-X
20. Khadzhynov D, Schmidt D, Hardt J, et al. The incidence of acute kidney injury and associated hospital mortality. *Dtsch Arztebl Int*. 2019;116(22):397-404. doi:10.3238/arztebl.2019.0397
21. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55(1):119-139. doi:10.1006/jcss.1997.1504
22. Xu J, Zhang X, Li W, Liu X, Han J. Joint multi-view 2D convolutional neural networks for 3D object classification. Bessiere C, ed. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization*; 2020:3202-3208. doi:10.24963/ijcai.2020/443
23. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263-1284. doi:10.1109/TKDE.2008.239
24. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin Pract*. 2012;120(4):c179-c184. doi:10.1159/000339789
25. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing knowledge consistently across health systems. *Yearb Med Inf*. 2017;26(01):139-147. doi:10.15265/IY-2017-018
26. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med*. 2018;46(7):1070-1077. doi:10.1097/CCM.0000000000003123
27. Cheng Z, Ye D, Zhu T, Zhou W, Yu PS, Zhu C. Multi-agent reinforcement learning via knowledge transfer with differentially private noise. *Int J Intell Syst*. 2022;37(1):799-828. doi:10.1002/int.22648

How to cite this article: Zhang X, Liu K, Yuan B, et al. A hybrid adaptive approach for instance transfer learning with dynamic and imbalanced data. *Int J Intell Syst*. 2022;1-18. doi:10.1002/int.23055