Journal of Biomedical Informatics 129 (2022) 104072

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

AutoScore-Imbalance: An interpretable machine learning tool for development of clinical scores with rare events data

Han Yuan [a], Feng Xie [a], Marcus Eng Hock Ong [a,b,c], Yilin Ning [a], Marcel Lucas Chee [d], Seyed Ehsan Saffari [a], Hairil Rizal Abdullah [a,e], Benjamin Alan Goldstein [a,f], Bibhas Chakraborty [a,f,g], Nan Liu [a,c,h,*]

[a] Duke-NUS Medical School, National University of Singapore, Singapore
[b] Department of Emergency Medicine, Singapore General Hospital, Singapore
[c] Health Services Research Centre, Singapore Health Services, Singapore
[d] Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia
[e] Department of Anaesthesiology, Singapore General Hospital, Singapore
[f] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States
[g] Department of Statistics and Data Science, National University of Singapore, Singapore
[h] Institute of Data Science, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Interpretable machine learning
Clinical score
Data imbalance
Automated machine learning
Electronic health records
Rare events

ABSTRACT

Background: Medical decision-making impacts both individual and public health. Clinical scores are commonly used among various decision-making models to determine the degree of disease deterioration at the bedside. AutoScore was proposed as a useful clinical score generator based on machine learning and a generalized linear model. However, its current framework still leaves room for improvement when addressing unbalanced data of rare events.

Methods: Using machine intelligence approaches, we developed AutoScore-Imbalance, which comprises three components: training dataset optimization, sample weight optimization, and adjusted AutoScore. Baseline techniques for performance comparison included the original AutoScore, full logistic regression, stepwise logistic regression, least absolute shrinkage and selection operator (LASSO), full random forest, and random forest with a reduced number of variables. These models were evaluated based on their area under the curve (AUC) in the receiver operating characteristic analysis and balanced accuracy (i.e., mean value of sensitivity and specificity). By utilizing a publicly accessible dataset from Beth Israel Deaconess Medical Center, we assessed the proposed model and baseline approaches to predict inpatient mortality.

Results: AutoScore-Imbalance outperformed baselines in terms of AUC and balanced accuracy. The nine-variable AutoScore-Imbalance sub-model achieved the highest AUC of 0.786 (0.732–0.839), while the eleven-variable original AutoScore obtained an AUC of 0.723 (0.663–0.783), and the logistic regression with 21 variables obtained an AUC of 0.743 (0.685–0.801). The AutoScore-Imbalance sub-model (using a down-sampling algorithm) yielded an AUC of 0.771 (0.718–0.823) with only five variables, demonstrating a good balance between performance and variable sparsity. Furthermore, AutoScore-Imbalance obtained the highest balanced accuracy of 0.757 (0.702–0.805), compared to 0.698 (0.643–0.753) by the original AutoScore and the maximum of 0.720 (0.664–0.769) by other baseline models.

Conclusions: We have developed an interpretable tool to handle clinical data imbalance, presented its structure, and demonstrated its superiority over baselines. The AutoScore-Imbalance tool can be applied to highly unbalanced datasets to gain further insight into rare medical events and facilitate real-world clinical decision-making.

* Corresponding author at: Duke-NUS Medical School, National University of Singapore, 8 College Road, Singapore 169857, Singapore.
E-mail address: liu.nan@duke-nus.edu.sg (N. Liu).

https://doi.org/10.1016/j.jbi.2022.104072
Received 14 November 2021; Received in revised form 10 March 2022; Accepted 7 April 2022
Available online 11 April 2022

1. Introduction

In medicine, decision-making encompasses diagnosis, treatment, disease prediction, and everyday conditions that impact individual and public health [1]. The increasing collection of clinical data, such as the growing electronic health records (EHRs) in hospitals [2] and advances in automated machine learning [3], have facilitated automatic medical decision-making. In general, clinicians prefer transparent and interpretable “glass box” models to complex “black box” models, such as artificial neural networks (ANNs) [4]. Interpretable models can be explained or presented in an understandable manner to a human being [5]. SHapley Additive exPlanations (SHAP) is an innovative approach to providing interpretability to “black box” models [6]. Likewise, local interpretable model-agnostic explanations (LIME) is another technique that explains classifiers’ predictions based on learning interpretable models around “black box” predictions [7]. However, both SHAP and LIME can be viewed as post hoc explanation methods that are not transparent enough for clinicians, who are more inclined to inherently transparent models like logistic regression [4].

Therefore, generic clinical scores are widely accepted and used by clinicians and nurses in hospitals for their transparency and accessibility [8,9]. Such scores take advantage of integer score points and categorize variables to identify clinical outcomes, trigger better care, and improve prognoses [10]. Clinical scores are often derived from expert consensus or through cohort analyses using traditional statistical methods [11].

To aid the development and validation of interpretable clinical scores, Xie et al. [12] proposed AutoScore, an automatic clinical score generator integrating machine learning and point-based score to ensure high discriminability and accessibility. The AutoScore comprises 6 modules, which include variable ranking, variable transformation, score derivation, model selection, score fine-tuning, and model evaluation. The AutoScore framework begins with the selection of top-ranking variables using machine learning. Then variable transformation converts continuous variables into categorical variables for the modeling of nonlinear effects. Next, the score derivation process creates a clinical score based on logistic regression. The model selection and score fine-tuning modules allow the user to determine the variables to include in the final score and the cutoff values to categorize continuous variables, respectively. Finally, the performance of the score is evaluated on an unseen test dataset.

Despite AutoScore’s ability to generate clinical score systems for a wide range of medical applications [13], it may not perform well when datasets are unbalanced due to the low prevalence of outcomes, i.e., rare medical events. Unbalanced datasets often make predictive models unreliable since they tend to focus on the dominant class and disregard the rare one [14]. Consequently, data imbalance may lead to poor prediction capabilities [15,16]. As per previous research, the model prediction capabilities were evaluated from two perspectives: evaluation metrics such as the area under the curve and the number of variables [17]. Therefore, we prefer models with fewer variables while achieving comparable or even better evaluation metrics.

Data imbalance has been addressed through a variety of approaches. Researchers traditionally used up-sampling (over-sampling) of minority samples, down-sampling (under-sampling) of majority samples, and sample weight adjustment to ensure a balanced distribution of classes [18]. The synthetic minority over-sampling technique (SMOTE) is a popular algorithm dealing with unbalanced datasets, which synthesizes new minority samples from several closest neighbors of the real minority samples [19]. In the medical domain, the unbalanced nature of many datasets has prompted researchers to propose novel approaches. Rahman et al. developed a cluster-based under-sampling technique [20], while Khalilia et al. sampled data to subgroups to build multiple random forest models and then ensembled these models to deal with data imbalance [21]. Li et al. introduced a Gaussian type fuzzy membership function for data down-sampling [22]. In recent years, generative adversarial networks (GANs) have been widely used in medical image syntheses to create new images for model development [23]. The GAN technique can also be applied to the generation of structured data, augmenting the proportion of minority samples in datasets and further improving the sample category distribution [24].

In this study, we sought to integrate methodologies for handling data imbalance into AutoScore and create an automated framework that allows for reliable risk scores to be derived even from unbalanced datasets. The AutoScore-Imbalance framework that we propose utilizes two novel components before the variable ranking in AutoScore to balance the training set using machine learning and to identify optimal sample weights for the rare class. The balanced training samples and sampling weights were then incorporated into the modified score derivation block, which employed weighted logistic regression to direct more attention towards the rare class, thereby reducing the likelihood of bias in the final scoring model.

2. Methods

2.1. *The AutoScore framework*

AutoScore [12] is a machine learning-based clinical score generator with six modules. Module 1 uses a random forest to rank variables according to their importance. Module 2 transforms variables by categorizing continuous variables to improve interpretation and cope with nonlinearity. Module 3 assigns scores to each categorized variable based on a logistic regression model. Module 4 determines the number of variables to include in the clinical score model depending on the tradeoff between model complexity and predictive performance. Module 5 incorporates clinical knowledge, in which cutoff points can be adjusted for the categorization of continuous variables. Lastly, Module 6 evaluates the performance of the score in an independent test dataset. The AutoScore framework provides a systematic and automated approach to the rapid development of a clinical score system, combining the advantage of machine learning in its discriminability and the strength of point-based score in its interpretability.

2.2. *Proposed AutoScore-Imbalance framework*

To deal with data imbalance and automate the development of sparse clinical scores, we proposed AutoScore-Imbalance, a novel extension to the original AutoScore framework. AutoScore-Imbalance adopts a nested structure to combine and reorganize AutoScore and its individual modules. It is comprised of three blocks, including newly introduced Block A (training data optimization) and Block B (sample weights optimization) to handle data imbalance, and Block C for final score derivation and evaluation based on the balanced data. The AutoScore-Imbalance framework is illustrated in Fig. 1. Using resampling and data synthesis techniques, Block A adjusts the raw unbalanced training dataset. Block B is designed to optimize sample (observation) weights, which are tuned to correct any imperfections that may lead to bias in the class proportion [25]. Block C comprises Modules 2 to 6 of the original AutoScore workflow, but now uses the relatively balanced datasets obtained in Block A to train the model, and uses a weighted logistic regression model in Module 3 (instead of the unweighted logistic regression in the original AutoScore) to incorporate the sample weights acquired from Block B.

2.2.1. *Block A: Training data optimization*

Using the unbalanced training dataset as input, Block A manipulates the data to produce a reasonably balanced dataset. The number of variables in the clinical score system is used as a hyperparameter in Block A and Block B for intermediate evaluations. As with random forest, we set this hyperparameter as the square root of the total number of variables [26].

Similar to AutoScore, AutoScore-Imbalance divides a full dataset into three parts: training data D, validation data, and test data. The training

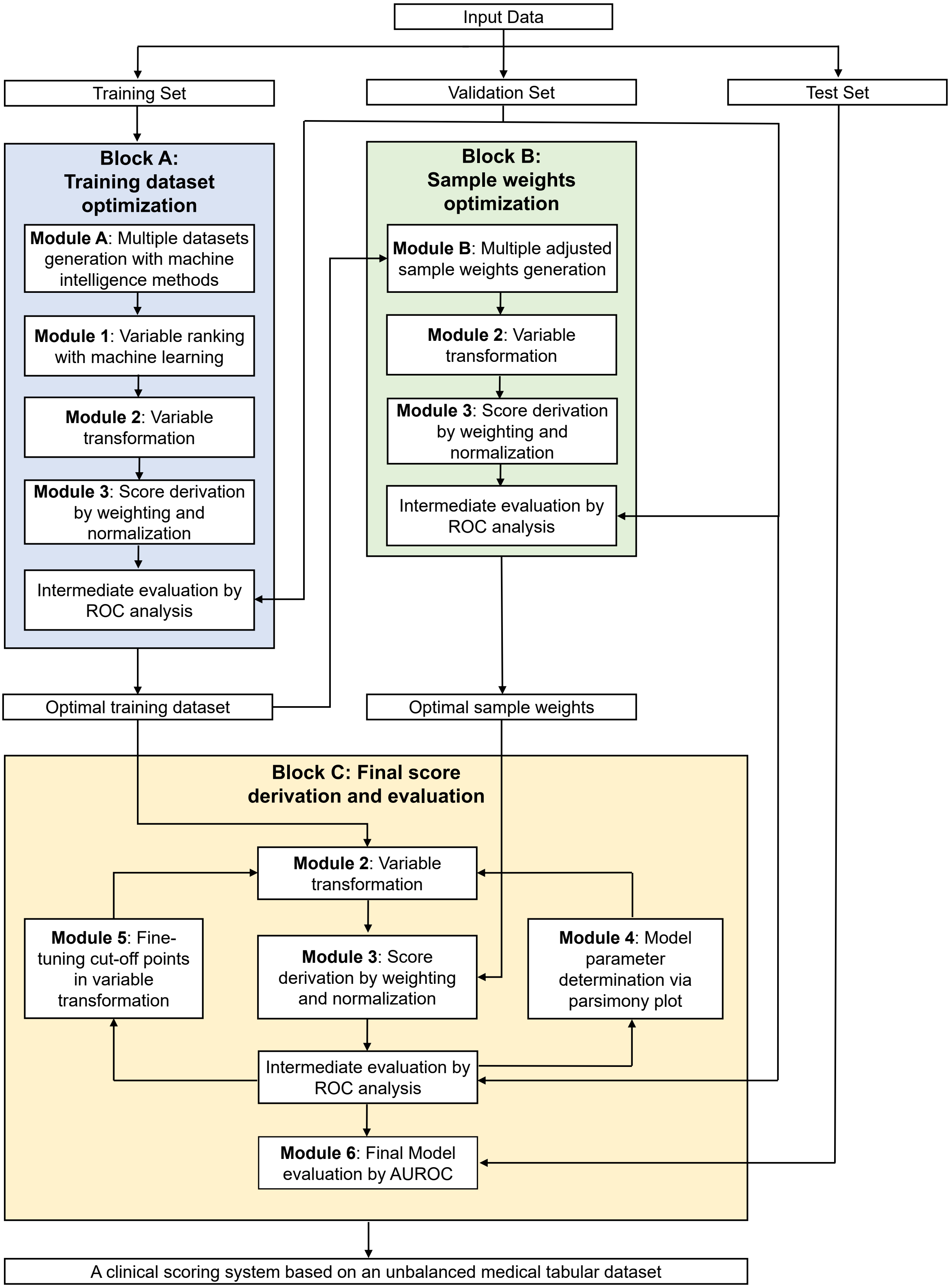

Input Data
Training Set
Validation Set
Test Set
Block A:
Training dataset
optimization
Module A: Multiple datasets generation with machine intelligence methods
Module 1: Variable ranking with machine learning
Module 2: Variable transformation
Module 3: Score derivation by weighting and normalization
Intermediate evaluation by ROC analysis
Block B:
Sample weights
optimization
Module B: Multiple adjusted sample weights generation
Module 2: Variable transformation
Module 3: Score derivation by weighting and normalization
Intermediate evaluation by ROC analysis
Optimal training dataset
Optimal sample weights
Block C: Final score derivation and evaluation
Module 2: Variable transformation
Module 5: Fine-tuning cut-off points in variable transformation
Module 3: Score derivation by weighting and normalization
Module 4: Model parameter determination via parsimony plot
Intermediate evaluation by ROC analysis
Module 6: Final Model evaluation by AUROC
A clinical scoring system based on an unbalanced medical tabular dataset

Fig. 1. Flowchart of the AutoScore-Imbalance framework.

dataset D is used to derive scores, the validation dataset is used for intermediate evaluation and parameter optimization, and the test dataset is reserved as unseen data for model performance assessment. The data-balancing methods in Block A are only applied to training data. The training dataset D of N samples is defined as follows,

$$D = \{D_i\}, i = 1, 2, 3, \cdots, N-1, N \tag{1}$$

where i represents the i th subject. D contains N_p minority samples (positive samples) and N_n majority samples (negative samples). In this study, we assume that rare clinical events are positive outcomes. The minority prevalence rate P is defined as N_p/N. As the first module in Block A, Module A aims to increase the minority rate from P to $P^{'}$ ($P < P^{'} \leq 0.5$) in the manipulated training dataset D'. By using up-sampling or other data augmentation techniques, the corresponding number of minority samples $N_p^{'}$ in the processed dataset is:

$$N_p^{'} = round\left(\frac{N_n}{1-P^{'}} - N_n\right) \tag{2}$$

When we use down-sampling to reduce sample size, the number of majority samples $N_n^{'}$ will become:

$$N_n^{'} = round\left(\frac{N_p}{P^{'}} - N_p\right) \tag{3}$$

In this study, we use an integer $\alpha > 0$ to denote the up-sampling ratio, which corresponds to the quotient in a Euclidean division of $N_p^{'}$ by N_p:

$$N_p^{'} = \alpha N_p + r \tag{4}$$

where r is the integer remainder with $0 \leq r < N_p$. In up-sampling operation, $\alpha - 1$ stands for the replication times of N_p and is constrained through $P^{'}$ and $N_p^{'}$ to avoid over-augmenting minor samples that the model over-fits the minority samples. r is the number of additional samples drawn from N_p data. For example, if we increase the minority sample size from $N_p = 100$ to $N_p^{'} = 205$, α is 2 and r is 5. The up-sampled dataset by Module A will include $2{\cdot}N_p$ minority samples (the original N_p and the duplicated N_p samples), 5 randomly selected minority samples from the original N_p data, and N_n majority samples. In SMOTE-based data augmentation, the number of synthesized samples is $(\alpha - 1){\cdot}N_p + r$, which is 105 based on the former example. Since SMOTE cannot generate non-integer times of artificial data, we achieve data augmentation in two steps: first, we use SMOTE to create $(\alpha - 1){\cdot}N_p$ synthetic samples from the original N_p minority data, which is 100 based on the former example; second, we randomly select 5 samples from the original N_p samples and then apply SMOTE to produce 5 synthetic ones from them.

There are three types of methods for training data optimization in Module A: resampling methods, data synthesis methods, and hybrid methods that combine both resampling and data synthesis strategies. Resampling methods include up-sampling of minority samples and down-sampling of majority samples. Data synthesis methods include SMOTE and GAN. These methods and their hybrid versions are used in Module A to generate a variety of processed datasets with different minority rates $P^{'}$ ($P < P^{'} \leq 0.5$). Subsequently, each processed dataset goes through Modules 1, 2, and 3 to construct risk scores using different numbers of top-ranking variables. These scores are evaluated based on the area under the curve (AUC) in the receiver operating characteristic (ROC) analysis applied to the validation dataset, and the optimally processed dataset will be returned. Next, we describe the various methods included in Module A, as shown in Table 1.

Resampling Methods: Simple data resampling techniques include up-sampling of minority samples and down-sampling of majority samples. They are the most commonly used techniques for dealing with data imbalance [27].

Table 1
List of methods and their detailed algorithms used in Module A.

Type	Method	Algorithm
Resampling Methods	Down-sampling	$D^{'} = \{D_1^{'}, D_2^{'}\}$, where $D_1^{'}$ is D_p, and $D_2^{'}$ is a set of $N_n^{'}$ selected samples without replacement from D_n
	Up-sampling	$D^{'} = \{D_1^{'}, D_2^{'}, D_3^{'}\}$, where $D_1^{'}$ is D, $D_2^{'}$ is $(\alpha - 1)$ times replication of D_p, and $D_3^{'}$ is r selected samples without replacement from D_p
Data Synthesis Methods	SMOTE	$D^{'} = \{D_1^{'}, D_2^{'}, D_3^{'}\}$, where $D_1^{'}$ is D, $D_2^{'}$ is a set of $(\alpha - 1){\cdot}N_p$ synthetic samples obtained from D_p through SMOTE, and $D_3^{'}$ is a set of r synthetic samples obtained from r randomly selected samples of D_p through SMOTE
	GAN	$D^{'} = \{D_1^{'}, D_2^{'}\}$, where $D_1^{'}$ is D, and $D_2^{'}$ is a set of $(N_p^{'} - N_p)$ synthetic minority samples obtained from D through GAN
Hybrid Methods	Up-sampling + Down-sampling	$D_h^{'}$ is an intermediate dataset with minority rate of $(P + P^{'})/2$ generated from D through up-sampling, and the final dataset $D^{'}$ with minority rate of $P^{'}$ is created from $D_h^{'}$ through down-sampling
	SMOTE + Down-sampling	$D_h^{'}$ is an intermediate dataset with minority rate of $(P + P^{'})/2$ generated from D through SMOTE, and the final dataset $D^{'}$ with minority rate of $P^{'}$ is created from $D_h^{'}$ through down-sampling
	GAN + Down-sampling	$D_h^{'}$ is an intermediate dataset with minority rate of $(P + P^{'})/2$ generated from D through GAN, and the final dataset $D^{'}$ with minority rate of $P^{'}$ is created from $D_h^{'}$ through down-sampling

D: The original training dataset.
$D^{'}$: The training datasets after Module A processing.
D_p: The minority samples in D.
D_n: The majority samples in D.
$D_1^{'}$, $D_2^{'}$, $D_3^{'}$: The first, second, and third part in $D^{'}$, individually.
$D_h^{'}$: The intermediate dataset in hybrid methods.
N: The total sample size in D.
N_p: The minority sample size in D.
N_n: The majority sample size in D.
$N_p^{'}$: The minority sample size in $D^{'}$ (See Equation (2)).
$N_n^{'}$: The majority sample size in $D^{'}$ (See Equation (3)).
P: The minority rate in D.
$P^{'}$: The minority rate in $D^{'}$.
α: the quotient in the Euclidean division of $N_p^{'}$ by N_p (See Equation (4)).
r: The remainder part of the Euclidean division of $N_p^{'}$ by N_p (See Equation (4)).

Data Synthesis Methods: SMOTE creates realistic "pseudo" minority samples through the following steps [19]: (i) Select Z nearest neighbors of each minority sample; (ii) Calculate differences between the sample and its Z nearest neighbors; (iii) Multiply differences by random numbers L $(0 < L < 1)$; (iv) Add those products to the sample to generate synthetic samples. By default, SMOTE creates Z (Z is an integer) sets of minority samples, i.e., $Z{\cdot}N_p$ samples, but cannot directly synthesize $Z{\cdot}N_p + \text{C}$ $(0 < \text{C} < N_p)$ artificial samples. Therefore, we customized the original SMOTE DMwR package [28] to fit our needs. GAN was initially proposed by Goodfellow et al. to generate synthetic images [23], and Xu et al. extended its application to structured data generation [29]. GAN has two adversarial components: a generative model and a discriminative model. Through iterative learning, the pseudo data generated by the generative model becomes increasingly similar to the real data. This study uses GAN to generate synthetic minority samples in the training dataset.

Hybrid Methods: Additionally, we explore several hybrid techniques that combine resampling and data synthesis methods. We first increase the minority sample quantity to an intermediate level by up-sampling, SMOTE, or GAN. Next, we reduce the sample size from the majority class to a specified level through down-sampling. Table 1 shows three hybrid methods used in our demonstration: up-sampling +

down-sampling, SMOTE + down-sampling, and GAN + down-sampling.

2.2.2. Block B: Sample weights optimization

Block B is designed to derive optimal sample weights for the majority and minority samples generated from Block A. The sample weight is defined as the contribution of each subject D_i to the loss function. In conventional logistic regression analysis, every subject contributes equally, as indicated by a common sample weight of one. When working with imbalanced data, where predictive models tend to be dominated by majority samples [14], we increase the weight assigned to minority samples to make erroneous predictions for this class more costly. With the sample weights, our proposed model no longer solely concentrates on the majority samples and ignores the minority samples [30]. To find an optimal weight for majority and minority samples, we developed the following approach. The sample weight of the majority sample is always set to one to ensure that the weight optimization process starts with treating each majority and minority sample equally. The procedure then gradually increases the minority sample's weight so that the loss function focuses more on the minority samples. To be specific, Block B receives the optimal dataset from Block A and outputs the optimal sample weights for samples in the optimal dataset. We use the AUC on the validation dataset as the criterion to select the optimal sample weights for minority samples in a grid search ranging from 1 to $w_{max} = N'_n/N'_p$ (to be rounded up to an integer) with a pre-set integer step, s.

2.2.3. Block C: Final score derivation and evaluation

Using the relatively balanced training dataset obtained from Block A and the optimal sample weights obtained from Block B as the inputs, Block C employs the original AutoScore framework (but beginning from Module 2) to generate sparse clinical scores. Module 2 converts continuous variables in the relatively balanced, restructured training dataset into categorical variables. In contrast to the original Module 3 in AutoScore, which uses an unweighted logistic regression model, Module 3 in Block C applies a weighted logistic regression model to the processed training dataset using the sample weights obtained from Block B. Module 4 determines the number of variables to be included in the clinical score using a parsimony plot, where top-ranked variables are chosen when there is no substantial improvement in AUC values with the addition of more variables. Module 5 allows users to customize cutoffs based on their domain knowledge. Module 6 evaluates the derived clinical score based on multiple performance evaluation metrics. Overall, Block C, the last step of the AutoScore-Imbalance framework, produces a standard clinical score table used for subsequent risk prediction.

2.3. Experiments

We demonstrated our AutoScore-Imbalance algorithm using the de-identified intensive care unit (ICU) dataset as our previous paper [12]. This dataset includes 21 continuous variables and 44,918 ICU admission episodes (including 3,958 positive episodes, defined as deaths that occurred during the hospital stay) of the Beth Israel Deaconess Medical Center between 2001 and 2012 (MIMIC-III dataset) [31]. We intend to create an unbalanced dataset for demonstrating our AutoScore-Imbalance method. There is currently no consensus on the definition of rare event rate. Several studies have reported an imbalance level (i.e., minority rate) of 1% [32], 0.5% [33], and 0.1% [34]. Our research adhered to the recommendation in a survey paper [35] that the minority-majority class rate is often less than or equal to 1% in a scenario of high-class imbalance. In this study, we randomly selected 404 positive (Death) and 40,000 negative (Survival) admission episodes to create a dataset with a 1% positive rate to demonstrate our methods under highly imbalanced conditions [36]. Additional experiments on datasets with positive rates of 0.5% and 0.1% were presented in the Appendix. To derive and evaluate the clinical score models, we split the entire dataset into three parts: training dataset (60%), validation dataset (20%), and test dataset (20%).

We compared the clinical score model derived by AutoScore-Imbalance with that of the original AutoScore, full logistic regression, stepwise logistic regression, LASSO, full random forest, and random forest with a reduced number of variables. We chose the "optimal" thresholds as the points nearest to the upper-left corner in the ROC curves to calculate performance metrics. In addition to commonly used metrics (AUC, sensitivity, specificity, negative predictive value [NPV], positive predictive value [PPV]), we used balanced accuracy (i.e., the average of the sensitivity and specificity values) to evaluate various model predictions on unbalanced datasets [37]. Sensitivity, specificity, NPV, PPV, and balanced accuracy were calculated with each model's optimal threshold, and their corresponding 95% confidence intervals (CIs) were obtained via bootstrapping [38].

2.4. Code Availability

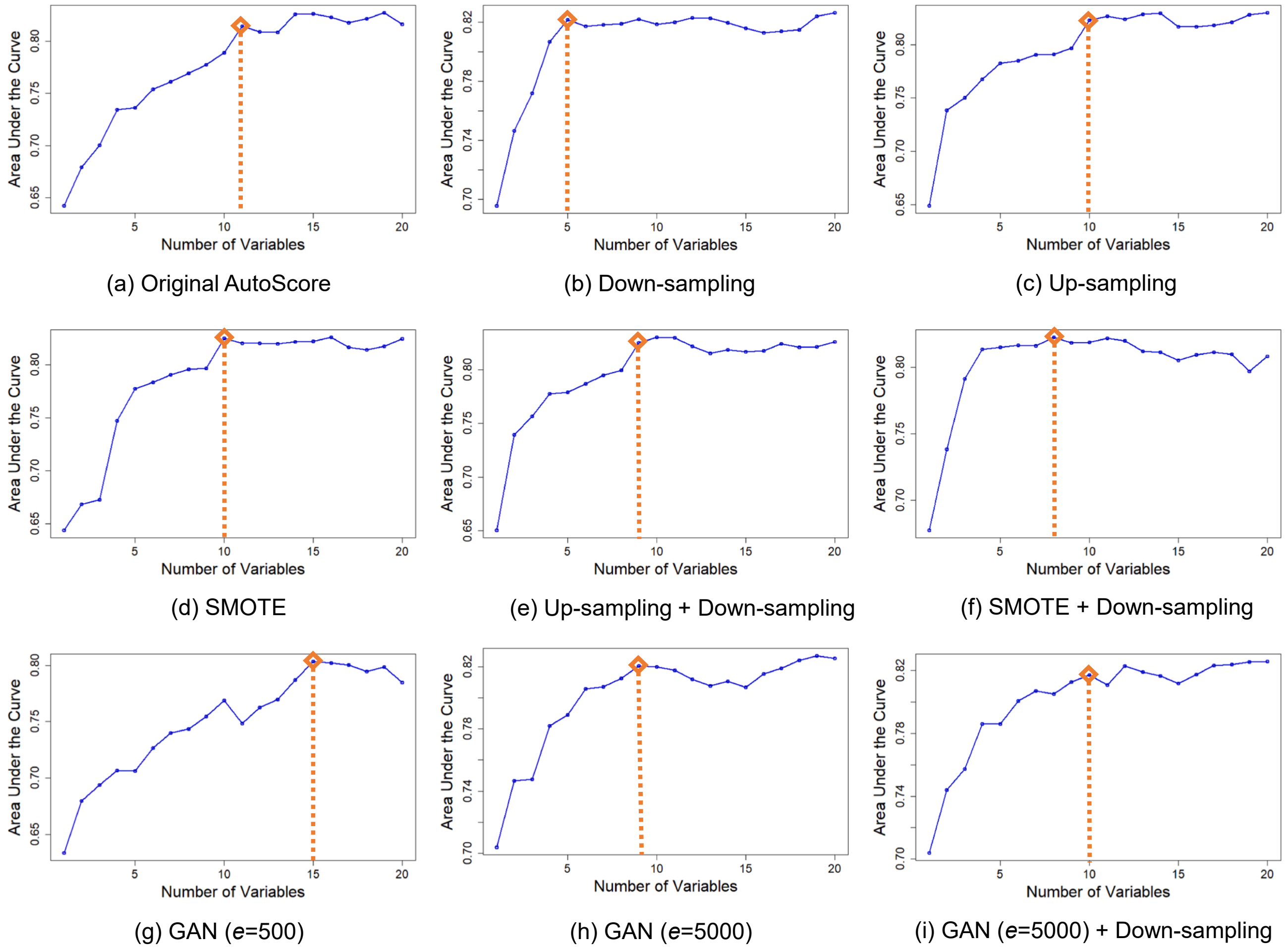
We implemented AutoScore-Imbalance in R software based on the AutoScore package [39] and made the corresponding software package available on GitHub [40].

3. Results

This study analyzed 40,404 ICU admission episodes. Specifically, the training dataset consisted of 24,244 episodes (60%, containing 244 positive samples), while the validation dataset consisted of 8,080 episodes (20%, including 80 positive samples), and the test dataset comprised 8,080 episodes (20%, containing 80 positive samples). We created a total of nine clinical score models, including AutoScore. Fig. 2 illustrates the parsimony plots for the original AutoScore and eight sub-models of the AutoScore-Imbalance framework. We manually selected the "near-optimal" number of variables to ensure that the performance could not be significantly improved by including additional variables in the models. Compared with the original AutoScore, most AutoScore-Imbalance sub-models achieved the "near-optimal" solutions with fewer variables, except for its sub-model (GAN, $e = 500$), which was insufficiently trained due to a small number of epochs in deep learning frameworks. Furthermore, the results in Table 2 show an enhanced performance of GAN-based clinical score models with increased training epochs.

We summarize the performance of different clinical score models in Table 2. The original AutoScore included 11 variables and achieved an AUC of 0.723 (95% CI: 0.663–0.783) and a balanced accuracy of 0.698 (95% CI: 0.643–0.753). Despite using fewer variables (apart from GAN [$e = 500$]), all sub-models of AutoScore-Imbalance had higher AUC and balanced accuracy values than the original AutoScore. Notably, AutoScore-Imbalance (down-sampling) achieved an AUC of 0.771 (95% CI: 0.718–0.823) and a balanced accuracy of 0.705 (95% CI: 0.651–0.759) with only five variables—less than half of the variables required by the original AutoScore. Furthermore, AutoScore-Imbalance yielded the highest balanced accuracy of 0.757 (95% CI: 0.702–0.805) using up-sampling and ten variables, higher than the value of 0.698 (95% CI: 0.643–0.753) in the original AutoScore with 11 variables and the maximum of 0.720 (95% CI: 0.664–0.769) in other baseline models (i.e., logistic regression and random forest models). Three AutoScore-Imbalance sub-models (SMOTE, up-sampling, up-sampling + down-sampling) outperformed all baseline methods (including random forest using all 21 variables) in terms of AUC and balanced accuracy while having better interpretability in the form of a clinical score.

Table 3 presents the selected variables by the original AutoScore and eight sub-models of the AutoScore-Imbalance framework. The AutoScore-Imbalance sub-models, except for GAN [$e = 500$], utilized fewer variables than the model generated by the original AutoScore. While there were overlaps in selected variables between AutoScore and AutoScore-Imbalance (e.g., both included heart rate, age, lactate, and



(a) Original AutoScore

(b) Down-sampling

(c) Up-sampling

(d) SMOTE

(e) Up-sampling + Down-sampling

(f) SMOTE + Down-sampling

(g) GAN (*e*=500)

(h) GAN (*e*=5000)

(i) GAN (*e*=5000) + Down-sampling

Fig. 2. Parsimony plots of the original AutoScore and AutoScore-Imbalance sub-models on the validation datasets (the orange diamond indicates the number of variables selected for each model).

respiration rate), there were also notable differences. For example, blood urea nitrogen was selected by all AutoScore-Imbalance clinical score models but not by the original AutoScore. Also, sodium, mean arterial pressure, creatinine, chloride, anion gap were only selected by AutoScore-Imbalance sub-models. Variables not selected by AutoScore could potentially distinguish rare events from dominating major events.

With AutoScore-Imbalance, a clinical score table can be generated for direct application to clinical practice. This score ranging from 0 to 100 (scores larger than 100 would be automatically rounded to 100; the range could be adjusted based on clinical needs) is used to identify patients at risk of suffering from adverse events. The minimum score is 0, which stands for no risk, while the maximum score of 100 means the highest risk. As an example, Table 4 summarizes fine-tuned clinical score tables for inpatient mortality prediction based on the nine-variable AutoScore-Imbalance sub-model (up-sampling + down-sampling), five-variable AutoScore-Imbalance sub-model (down-sampling), and eleven-variable original AutoScore clinical score model.

4. Discussion

This study developed an interpretable machine learning model that coped with data imbalance and generated trustworthy clinical scores. Using an unbalanced real-world dataset, the proposed AutoScore-Imbalance framework achieved an improved prediction performance than the original AutoScore with fewer variables. Due to its sparsity, parsimony in variable selection, and intrinsically interpretable output format (See Table 4), a clinical score derived from AutoScore-Imbalance is practical for use at the bedside compared with both statistical methods and “black box” machine learning models.

When studying various approaches to handle data imbalance, conventional methods (up-sampling, down-sampling, and SMOTE) have proven to be effective in handling unbalanced datasets to develop risk scores. Additionally, AutoScore-Imbalance in conjunction with a modern technique, GAN, displayed comparable performance and outperformed the original AutoScore when training epochs were increased. GAN needs sufficient training epochs to generate reliable synthetic samples [41]. Therefore, a suitable training epoch for GAN should be determined upon assessment of a validation dataset.

Compared with AutoScore and AutoScore-Imbalance, baseline methods have an intrinsic advantage, as they can build clinical score models using high-resolution, continuous variables rather than categorized variables. This superiority, however, poses a limitation to the utility of baseline models in clinical risk prediction, where sparse and itemized scores are favored. In this regard, it is noteworthy that AutoScore-Imbalance produced improved prediction results while preserving score sparsity and clinical usability. Furthermore, among all baseline models, LASSO demonstrated the best prediction ability, which is in accordance with the fact that LASSO is effective in identifying important variables in unbalanced datasets [42].

The strength of AutoScore-Imbalance lies in its ability to address data imbalance while at the same time producing reliable and interpretable clinical scores. The outputs of AutoScore-Imbalance, integer-based, itemized clinical scores, are better received by clinicians than complicated “black box” models. For example, a prospective observational

Table 2
Performance of the original AutoScore, AutoScore-Imbalance, and baselines.

Models		m [a]	Threshold [b]	AUC [c]	Sensitivity [d]	Specificity [e]	Balanced Accuracy [f]	NPV [g]	PPV [h]
AutoScore		11	58	0.723 (0.663–0.783)	0.700 (0.600–0.800)	0.696 (0.686–0.706)	0.698 (0.643–0.753)	0.996 (0.994–0.997)	0.022 (0.019–0.026)
Full LR		21	0.007	0.743 (0.685–0.801)	0.787 (0.700–0.875)	0.602 (0.591–0.612)	0.695 (0.646–0.744)	0.996 (0.995–0.998)	0.019 (0.017–0.022)
Stepwise LR		16	0.011	0.737 (0.679–0.796)	0.637 (0.537–0.750)	0.748 (0.738–0.757)	0.693 (0.638–0.754)	0.995 (0.994–0.997)	0.025 (0.021–0.029)
LASSO		6	-4.586	0.768 (0.716–0.820)	0.738 (0.637–0.825)	0.702 (0.691–0.712)	0.720 (0.664–0.769)	0.996 (0.995–0.998)	0.024 (0.021–0.027)
Full RF		21	0.005	0.743 (0.685–0.800)	0.775 (0.675–0.863)	0.602 (0.591–0.613)	0.689 (0.633–0.738)	0.996 (0.995–0.998)	0.019 (0.017–0.021)
Parsimony RF		11	0.005	0.714 (0.655–0.772)	0.750 (0.650–0.838)	0.592 (0.581–0.602)	0.671 (0.616–0.720)	0.996 (0.994–0.997)	0.018 (0.016–0.020)
AutoScore-Imbalance	SMOTE	10	55	0.779 (0.727–0.832)	0.775 (0.675–0.863)	0.685 (0.675–0.695)	0.730 (0.675–0.779)	0.997 (0.995–0.998)	0.024 (0.021–0.027)
	US	10	57	0.780 (0.724–0.835)	0.738 (0.637–0.825)	0.776 (0.767–0.785)	0.757 (0.702–0.805)	0.997 (0.995–0.998)	0.032 (0.028–0.036)
	DS	5	68	0.771 (0.718–0.823)	0.537 (0.437–0.637)	0.873 (0.865–0.880)	0.705 (0.651–0.759)	0.995 (0.994–0.996)	0.041 (0.032–0.048)
	US + DS	9	55	0.786 (0.732–0.839)	0.725 (0.625–0.825)	0.758 (0.749–0.768)	0.742 (0.687–0.797)	0.996 (0.995–0.998)	0.029 (0.025–0.033)
	SMOTE + DS	8	55	0.767 (0.714–0.820)	0.750 (0.650–0.838)	0.689 (0.678–0.699)	0.720 (0.664–0.769)	0.996 (0.995–0.998)	0.024 (0.020–0.026)
	GAN (*e* = 500)	15	49	0.744 (0.686–0.802)	0.625 (0.525–0.725)	0.795 (0.786–0.803)	0.710 (0.656–0.764)	0.995 (0.994–0.997)	0.030 (0.024–0.035)
	GAN (*e* = 5000)	9	36	0.753 (0.704–0.803)	0.725 (0.625–0.825)	0.690 (0.679–0.700)	0.708 (0.652–0.763)	0.996 (0.995–0.997)	0.023 (0.020–0.026)
	GAN (*e* = 5000) + DS	10	36	0.759 (0.710–0.809)	0.738 (0.637–0.838)	0.699 (0.689–0.709)	0.719 (0.663–0.774)	0.996 (0.995–0.998)	0.024 (0.021–0.027)

LR: Logistic regression.
LASSO: Least absolute shrinkage and selection operator.
RF: Random forest.
SMOTE: Synthetic minority over-sampling technique.
US: Up-sampling.
DS: Down-sampling.
GAN: Generative adversarial networks.
e: Training epochs of GAN.
[a] The number of variables included in each model.
[b] Optimal cutoff values, defined as the points nearest to the upper-left corner in the ROC curves.
[c] AUC: the area under the ROC curve.
[d] Sensitivity = TP / (TP + FN), TP: true positive, FN: false negative.
[e] Specificity = TN / (TN + FP), TN: true negative, FP: false positive.
[f] Balanced Accuracy = (Sensitivity + Specificity)/2.
[g] NPV: negative predictive value = TN/ (TN + FN).
[h] PPV: positive predictive value = TP / (TP + FP).

study has shown that clinical scores improved patient safety in surgical wards and should be implemented in practice [43]. The variable ranking module of AutoScore-Imbalance offers a straightforward evaluation of variable importance to the outcome of interest, providing an objective support to clinicians' assessments based on their experience and domain knowledge. Besides being a convenient tool for risk assessment, the clinical score derived from AutoScore-Imbalance also quantifies the impact of variables on the outcome, thus enabling a transparent score interpretation via the points assigned to each variable category. As an example, the five-variable AutoScore-Imbalance (down-sampling) score model suggests that high blood urea nitrogen level (32.5 mg/L or higher) has a substantial impact on the risk of inpatient mortality, as indicated by the corresponding point of 33, whereas temperature only shows a marginal effect on the outcome (the maximum point associated is 5).

The novel AutoScore-Imbalance framework provides a practical and referable two-step pipeline for modifying training data and adjusting sample weights to handle data imbalance, which is not explicitly limited to clinical applications. Moreover, AutoScore-Imbalance is designed in a modular manner to make it easy to incorporate other state-of-the-art techniques for efficient score derivation and evaluation. When making high-stakes decisions within clinical settings, the use of AutoScore-Imbalance-derived clinical scores requires careful selection of candidate variables. In our example, all variables came from objective instruments rather than the experience or knowledge of the physicians; this is consistent with recent research showing that objective and data-driven clinical tools are capable of improving the triage process [44]. In the meantime, it is crucial to emphasize that prospective studies are still necessary to confirm the scores' sustainability, cost-effectiveness, and physician perceived acceptability [45].

The contributions in this work can lead to new lines in rare events-related decision makings. Several new questions emerge in light of the discoveries presented here. First, we evaluated AutoScore-Imbalance on a dataset created from EHR data with a minority rate of 1%. Additionally, we created datasets with similar sample sizes, but lower minority rates (0.5% and 0.1%) from the same EHR data to further examine AutoScore-Imbalance (see Appendix). Further testing of AutoScore-Imbalance will be conducted in future studies with varying minority rates and sample sizes. Second, we examined only one clinical application using the MIMIC-III database for demonstration. Given the low percentage of positive samples, only 80 positive cases were in the test set, which led to relatively low PPV values and overlapping CIs of

Table 3
Selected variables in different clinical scores.

Models	Original AutoScore	AutoScore-Imbalance							
		SMOTE	US	DS	US + DS	SMOTE + DS	GAN (e = 500)	GAN (e = 5000)	GAN (e = 5000) + DS
m [a]	11	10	10	5	9	8	15	9	10
Temperature (°C)	✓	✓	✓	✓	✓		✓		
Heart rate (beats/min)	✓	✓	✓		✓	✓	✓	✓	✓
Age (years)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Respiration rate (breaths/min)	✓	✓	✓	✓	✓	✓	✓		
Systolic blood pressure (mm Hg)	✓	✓	✓		✓	✓	✓		
SpO_2 (%)	✓	✓	✓		✓	✓	✓		✓
White blood cells (thousand per microliter)	✓	✓	✓		✓		✓		
Diastolic blood pressure (mm Hg)	✓						✓		
Platelet (thousand per microliter)	✓	✓	✓				✓		
Glucose (mg/dL)	✓								
Sodium (mmol/L)							✓	✓	✓
Lactate (mmol/L)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean arterial pressure (mm Hg)							✓		
Potassium (mmol/L)									
Bicarbonate (mmol/L)							✓	✓	✓
Blood urea nitrogen (mg/dL)		✓	✓	✓	✓	✓	✓	✓	✓
Hematocrit (%)									
Creatinine (μmol/L)								✓	✓
Hemoglobin (g/dL)									
Chloride (mEq/L)							✓	✓	✓
Anion gap (mEq/L)						✓		✓	✓

✓: The variable is selected in this model.
e: Training epochs of generative adversarial networks (GAN).
SMOTE: Synthetic minority over-sampling technique.
US: Up-sampling.
DS: Down-sampling.
GAN: Generative adversarial networks.
SpO_2: Peripheral capillary oxygen saturation.
[a] Parameter m is the number of variables included in the AutoScore model.

evaluation metrics for all methods. It is, therefore, necessary to perform additional validations under different settings. Third, the heterogeneity of clinical applications prevented us from recommending the best way to handle data imbalance. In general, SMOTE is a popular tool for augmenting data but might not be effective in specific scenarios, as with other sampling techniques [46]. And, it is worth considering GAN for dealing with high-dimensional data [47]. Lastly, our method was designed for tabular data without considering time-series data [48]. To develop a complete AutoScore-Imbalance solution, further studies will be required to extend its application to longitudinal data [49].

5. Conclusion

We proposed an interpretable machine learning-based AutoScore-Imbalance framework for automatic clinical score generation that addresses data imbalance. Compared with baseline models, this innovative framework presented a capability of developing good-performing and interpretable clinical scores on unbalanced datasets. We anticipate that this score generator will hold great potential in creating and evaluating sparse and itemized clinical scores in a variety of settings.

Funding

This study was supported by Duke-NUS Medical School, Singapore. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Table 4
A summary of the nine-variable AutoScore-Imbalance (up-sampling + down-sampling), five-variable AutoScore-Imbalance (down-sampling), and eleven-variable original AutoScore for inpatient mortality prediction.

AutoScore-Imbalance				Original AutoScore	
Up-sampling + Down-sampling		Down-sampling			
Variables and Interval [a]	Point	Variables and Interval	Point	Variables and Interval	Point
Age (years)					
< 50	0	< 50	0	< 50	0
50–65	7	50–65	13	50–65	9
65–75	13	65–75	18	65–75	14
≥ 75	17	≥ 75	22	≥ 75	19
Lactate (mmol/L)					
< 1.7	6	< 1.7	13	< 1.7	7
1.7–1.8	2	1.7–1.8	0	1.7–1.8	4
1.8–2	0	1.8–2.3	5	1.8–1.95	0
≥ 2	13	≥ 2.3	24	≥ 1.95	12
Temperature (°C)					
< 36.5	4	< 36.5	5	< 36.5	6
≥ 36.5	0	≥ 36.5	0	≥ 36.5	0
Heart rate (beats/min)					
< 74	5			< 74	5
74–84	0			74–84	0
84–95	3			84–95	4
≥ 95	16			≥ 95	16
SpO_2 [b] (%)					
< 96.3	3			< 96.3	4
96.3–98.8	0			96.3–97.6	1
≥ 98.8	6			97.6–98.7	0
				≥ 98.7	5
Systolic blood pressure (mm Hg)					
< 110	8			< 110	5
110–120	1			110–130	0
120–130	0			≥ 130	5
≥ 130	5				
White blood cells (thousand per microliter)					
< 7.95	6			< 7.9	5
7.95–10.7	0			7.9–10.6	0
≥ 10.7	12			10.6–14	11
				≥ 14	13
Respiration rate (breaths/min)					
< 16	2	< 16	2	< 16	3
16–18	0	16–18	0	16–18	0
18–21	3	18–21	4	18–20	5
≥ 21	7	≥ 21	16	≥ 20	6
Blood urea nitrogen (mg/dL)					
< 12.5	1	< 13.5	7		
12.5–18	0	13.5–19	0		
18–29	8	19–32.5	20		
≥ 29	17	≥ 32.5	33		
Platelet (thousand per microliter)					
				< 160	7
				160–210	4
				210–280	0
				≥ 280	3
Glucose (mg/dL)					
				< 111	4
				111–129	0
				129–153	4
				≥ 153	2
Diastolic blood pressure (mm Hg)					
				< 71.1	8
				71.1–77.2	6
				77.2–85.2	0
				≥ 85.2	5

[a] An interval (q1-q2) represents $q1 \leq x < q2$.
[b] SpO_2: peripheral capillary oxygen saturation.

Data Availability

We used de-identified critical care unit data from the Beth Israel Deaconess Medical Center between 2001 and 2012 (MIMIC-III dataset) [31], which is available at https://archive.physionet.org/physiobank/database/mimic3cdb/.

CRediT authorship contribution statement

Han Yuan: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Feng Xie:** Data curation, Investigation, Methodology, Validation, Writing – review & editing. **Marcus Eng Hock Ong:** Investigation, Methodology, Validation, Writing – review & editing. **Yilin Ning:** Investigation, Methodology, Validation, Writing – review & editing. **Marcel Lucas Chee:** Investigation, Methodology, Validation, Writing – review & editing. **Seyed Ehsan Saffari:** Investigation, Methodology, Validation, Writing – review & editing. **Hairil Rizal Abdullah:** Investigation, Methodology, Validation, Writing – review & editing. **Benjamin Alan Goldstein:** Investigation, Methodology, Validation, Writing – review & editing. **Bibhas Chakraborty:** Investigation, Methodology, Validation, Writing – review & editing. **Nan Liu:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104072.

References

[1] M. Li, G.B. Chapman, Medical decision making, in: R.H. Paul, L.E. Salminen, J. Heaps, L.M. Cohen (Eds.), The Wiley Encyclopedia of Health Psychology, Wiley, 2020, pp. 347–353.

[2] A.K. Jha, C.M. DesRoches, E.G. Campbell, K. Donelan, S.R. Rao, T.G. Ferris, A. Shields, S. Rosenbaum, D. Blumenthal, Use of electronic health records in US hospitals, N. Engl. J. Med. 360 (16) (2009) 1628–1638.

[3] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artif. Intell. Med. 104 (2020).

[4] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.

[5] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. 32 (24) (2020) 18069–18083.

[6] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[7] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016, 2016, pp. 1135–1144.

[8] G.B. Smith, D.R. Prytherch, P. Meredith, P.E. Schmidt, P.I. Featherstone, The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death, Resuscitation 84 (4) (2013) 465–470.

[9] M.M. Churpek, T.C. Yuen, S.Y. Park, D.O. Meltzer, J.B. Hall, D.P. Edelson, Derivation of a cardiac arrest prediction model using ward vital signs, Crit. Care Med. 40 (7) (2012) 2102–2108.

[10] S. Leteurtre, F. Leclerc, J. Wirth, O. Noizet, E. Magnenant, A. Sadik, C. Fourier, R. Cremer, Can generic paediatric mortality scores calculated 4 hours after admission be used as inclusion criteria for clinical trials? Crit. Care 8 (4) (2004) 1–9.

[11] J.P. Greving, M.J.H. Wermer, R.D. Brown, A. Morita, S. Juvela, M. Yonekura, T. Ishibashi, J.C. Torner, T. Nakayama, G.J.E. Rinkel, A. Algra, Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies, The Lancet Neurology 13 (1) (2014) 59–66.

[12] F. Xie, B. Chakraborty, M.E.H. Ong, B.A. Goldstein, N. Liu, AutoScore: A Machine Learning-Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records, JMIR medical informatics 8 (10) (2020), e21798.

[13] F. Xie, M.E.H. Ong, J.N.M.H. Liew, K.B.K. Tan, A.F.W. Ho, G.D. Nadarajan, L. L. Low, Y.H. Kwan, B.A. Goldstein, D.B. Matchar, B. Chakraborty, N. Liu, Development and Assessment of an Interpretable Machine Learning Triage Tool for

Estimating Mortality After Emergency Admissions, JAMA network open 4 (8) (2021) e2118467.

[14] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Min. Knowl. Disc. 28 (1) (2014) 92–122.

[15] A.J. Larrazabal, N. Nieto, V. Peterson, D.H. Milone, E. Ferrante, Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, Proc. Natl. Acad. Sci. 117 (23) (2020) 12592–12594.

[16] Y.X. Zhao, H. Yuan, Y. Wu, Prediction of Adverse Drug Reaction using Machine Learning and Deep Learning Based on an Imbalanced Electronic Medical Records Dataset, in: 5th International Conference on Medical and Health Informatics, ACM, 2021, pp. 17–21.

[17] N. Liu, Z.X. Koh, E.-P. Chua, L.-L. Tan, Z. Lin, B. Mirza, M.E.H. Ong, Risk scoring for prediction of acute cardiac complications from imbalanced clinical data, IEEE J. Biomed. Health. Inf. 18 (6) (2014) 1894–1902.

[18] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: A review, Int. J. Pattern Recognit Artif Intell. 23 (04) (2009) 687–719.

[19] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[20] M.M. Rahman, D.N. Davis, Addressing the class imbalance problem in medical datasets, International Journal of Machine Learning and Computing 3 (2) (2013) 224.

[21] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Med. Inf. Decis. Making 11 (1) (2011) 1–13.

[22] D.-C. Li, C.-W. Liu, S.C. Hu, A learning method for the class imbalance problem with medical data sets, Comput. Biol. Med. 40 (5) (2010) 509–518.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[24] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: An overview, IEEE Signal Process Mag. 35 (1) (2018) 53–65.

[25] W. Lee, C.-H. Jun, J.-S. Lee, Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification, Inf. Sci. 381 (2017) 92–103.

[26] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32.

[27] E. Rendon, R. Alejo, C. Castorena, F.J. Isidro-Ortega, E.E. Granda-Gutierrez, Data sampling methods to deal with the big data multi-class imbalance problem, Applied Sciences 10 (4) (2020) 1276.

[28] L. Torgo. Data Mining with R, learning with case studies, Chapman and Hall/CRC, 2010. http://www.liaad.up.pt/~ltorgo/DataMiningWithR.

[29] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, Advances in Neural Information Processing Systems 32 (2019).

[30] S. Jiang, X. Lu, WeSamBE: A weight-sample-based method for background subtraction, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2105–2115.

[31] A.E.W. Johnson, T.J. Pollard, L.u. Shen, L.-W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016).

[32] Yang YY, Akbarzadeh HA Khorshidi H, Aickelin UU, Nevgi AA, Ekinci EE: On the Importance of Diversity in Re-Sampling for Imbalanced Data and Rare Events in Mortality Risk Models. In: 2021 Australasian Computer Science Week Multiconference: 2021; 2021: 1-8.

[33] P.W. Lane, Meta-analysis of incidence of rare events, Stat. Methods Med. Res. 22 (2) (2013) 117–132.

[34] E.W. Chan, K.Q. Liu, C.S. Chui, C.W. Sing, L.Y. Wong, I.C. Wong, Adverse drug reactions–examples of detection of rare events using databases, Br. J. Clin. Pharmacol. 80 (4) (2015) 855–861.

[35] J.L. Leevy, T.M. Khoshgoftaar, R.A. Bauder, N. Seliya, A survey on addressing high-class imbalance in big data, Journal of Big Data 5 (1) (2018) 1–30.

[36] F. Thabtah, S. Hammoud, F. Kamalov, A. Gonsalves, Data imbalance in classification: Experimental evaluation, Inf. Sci. 513 (2020) 429–441.

[37] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th international conference on pattern recognition: 2010: IEEE, 2010, pp. 3121–3124.

[38] B. Efron, Bootstrap methods: another look at the jackknife, in: Breakthroughs in statistics, Springer, 1992, pp. 569–593.

[39] F. Xie, Y. Ning, H. Yuan, E. Saffari, B. Chakraborty, N. Liu, Package 'AutoScore': An Interpretable Machine Learning-Based Automatic Clinical Score Generator, R package version 0.2.0, 2021. Available from https://cran.r-project.org/package=AutoScore.

[40] H. Yuan, F. Xie, Y. Ning, N. Liu, Package 'AutoScore-Imbalance', 2022. Available from https://github.com/nliulab/AutoScore-Imbalance.

[41] Gruber T, Cammerer S, Hoydis J, ten Brink S: On deep learning-based channel decoding. In: 2017 51st Annual Conference on Information Sciences and Systems (CISS): 2017: IEEE; 2017: 1-6.

[42] N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4) (2010) 417–473.

[43] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, N. Keeling, The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study, The Annals of The Royal College of Surgeons of England 88 (6) (2006) 571–575.

[44] J. Miles, J. Turner, R. Jacques, J. Williams, S. Mason, Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review, Diagnostic and prognostic research 4 (1) (2020) 1–12.

[45] Z. Khadjesari, S. Boufkhed, S. Vitoratou, L. Schatte, A. Ziemann, C. Daskalopoulou, E. Uglik-Marucha, N. Sevdalis, L. Hull, Implementation outcome instruments for use in physical healthcare settings: a systematic review, Implementation Science 15 (1) (2020) 1–16.

[46] C. Pak, T.T. Wang, X.H. Su, An empirical study on software defect prediction using over-sampling by SMOTE, Int. J. Software Eng. Knowl. Eng. 28 (06) (2018) 811–830.

[47] C. Wang, H. Hu, Y. Lu, A solvable high-dimensional model of GAN, Advances in Neural Information Processing Systems 32 (2019).

[48] J.J. Zhang, Z. Sun, H. Yuan, M. Wang, Alternatives to the Kaplan-Meier estimator of progression-free survival, The International Journal of Biostatistics 17 (1) (2021) 99–115.

[49] F. Xie, Y. Ning, H. Yuan, B.A. Goldstein, M.E.H. Ong, N. Liu, B. Chakraborty, AutoScore-Survival: Developing interpretable machine learning-based time-to-event scores with right-censored survival data, J. Biomed. Inform. 125 (2022), 103959.