Journal Pre-proof

A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset

Tianyu Liu, Wenhui Fan, Cheng Wu

PII: S0933-3657(19)30229-5

DOI: https://doi.org/10.1016/j.artmed.2019.101723

Reference: ARTMED 101723

To appear in: *Artificial Intelligence In Medicine*

Received Date: 28 March 2019

Revised Date: 12 August 2019

Accepted Date: 6 September 2019

Please cite this article as: { doi: https://doi.org/

A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset

Tianyu Liu[a,1], Wenhui Fan[a,*], Cheng Wu[a,1]

[a]*Department of Automation, Tsinghua University, Beijing, China*

Abstract

Background and Objective : Cerebral stroke has become a significant global public health issue in recent years. The ideal solution to this concern is to prevent in advance by controlling related metabolic factors. However, it is difficult for medical staff to decide whether special precautions are needed for a potential patient only based on the monitoring of physiological indicators unless they are obviously abnormal. This paper will develop a hybrid machine learning approach to predict cerebral stroke for clinical diagnosis based on the physiological data with incompleteness and class imbalance.

Methods: Two steps are involved in the whole process. Firstly, random forest regression is adopted to impute missing values before classification. Secondly, an automated hyperparameter optimization(AutoHPO) based on deep neural network(DNN) is applied to stroke prediction on an imbalanced dataset.

Results: The medical dataset contains 43400 records of potential patients which includes 783 occurrences of stroke. The false negative rate from our prediction approach is only 19.1%, which has reduced by an average of 51.5% in comparison to other traditional approaches. The false positive rate, accuracy and sensitivity predicted by the proposed approach are respectively 33.1%, 71.6%, 67.4%.

Conclusion: The approach proposed in this paper has effectively reduced the false negative rate with a relatively high overall accuracy, which means a successful decrease in the misdiagnosis rate for stroke prediction. The results are

*Corresponding author
Email address: fanwenhui@tsinghua.edu.cn (Wenhui Fan)

more reliable and valid as the reference in stroke prognosis, and also can be acquired conveniently at a low cost.

Keywords: Stroke prediction, Clinical decision, Class imbalance, Hybrid machine learning, AutoHPO

1. Introduction

Cerebral stroke, a disease with severe morbidity, disability and mortality, has become one of the major threats to public health worldwide. According to the research of GBD[1], disability adjusted of life years (DALYs) caused by stroke rank secondly only after the ischemic heart disease, and the details are shown as Fig.1[1, 2]. Although the pathogenesis of stroke is still not quite clear, it is generally acknowledged that stroke is closely related to abnormal metabolic indicators for both hemorrhagic stroke and ischemic stroke[3]. Given the fact that over 90% of metabolic risk factors of this disease can be controllable, more attention should be attached to the prevention[4]. In this paper, we attempt to predict potential stroke attacks based on related metabolic indicators. The objective is to decide that whether preventive intervention is necessary in advance and provide some diagnostic support for clinical medicine.

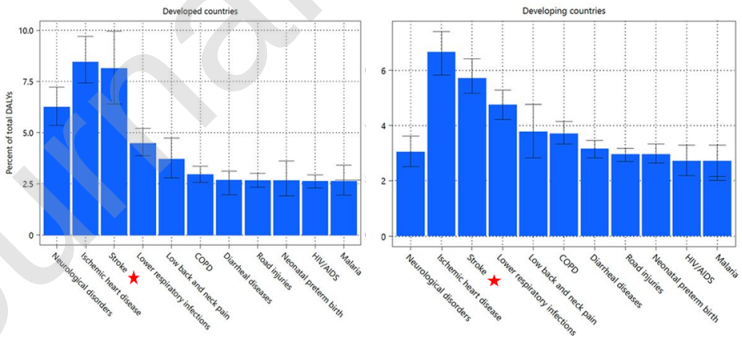


Fig. 1. Proportion (%) of leading causes of DALYs in developed and developing countries from 1990 to 2013.

[1]Global Burden of Diseases, Injuries and Risk Factors Study, in 2017

1.1. Opportunities and challenges

Medical dataset usually consists of patient symptoms and health conditions. In recent years, machine learning has served as an advanced diagnosis and prognosis technique, which can classify medical data into predefined class labels, such as sick or non-sick[5]. Such a technique has provided the opportunity to successfully implement machine learning for higher accuracy and efficiency in clinical diagnonsis[6, 7, 8]. However, it still has manifold limitations and vulnerability from the application perspective. Data completeness and class balance are difficult to be satisfied in reality, which brings great challenges to the performance of the prediction model. It must be realized that a high imbalance ratio of the data could even lead to meaningless prediction.

1.2. Related work

An increasing number of researches have investigated the application of machine learning models in stroke prediction in the last decade[9, 10]. For example, Khosla et al.[11] successfully predicted the outcome of stroke based on the Cox proportional hazards model and its better performance was also verified through comparison with other popular methods. Then, an evolving spiking neural network reservoir system (eSNNr) was developed to reduce the prediction error[12]. Arslan et al.[13] assessed various learning-based methods to predict stroke, among which support vector machine (SVM) and stochastic gradient boosting (SGB) methods are regarded as the remarkable ones.

Most of the existing researches about stroke prediction are concerned with the complete and class balance dataset, but few medical datasets can strictly meet such requirements. For the incomplete data, a missing value imputation method based on iterative mechanism has shown an acceptable prediction accuracy[14, 15]. With further development of research, Zhang et al.[16] proposed the expectation maximization algorithms based Bayesian classification to impute missing values, while class imbalance is still an outstanding problem. In the actual clinical practice, the stroke dataset suffers from class imbalance in nature. Thus it is necessary to consider class imbalance if we utilize learning-based

methods to predict stroke.

Class imbalance means imbalanced sample distribution among classes where the rare classes are called minority classes, and the prevalent ones are majority classes[17]. This issue has been relieved from the data level and algorithm level. From the data's perspective, re-sampling is a main approach to rebalance data distribution by under-sampling or over-sampling, where the selection of these two methods is decided by the size of minority class in the original dataset[18, 19]. However, random sampling inevitably results in implicit information loss or overlap. To this issue, there have been a number of attempts at improving the sampling efficiency and accuracy, such as distance-based sampling[20], SMOTE[21] and cluster sampling[22]. From the perspective of algorithm design, modifications on loss function and objective function in the traditional prediction models are widely applied[23, 24, 25]. One of the strategies is to redefine the classification boundary with the linearly separable dataset[26]. For classification on nonlinear and nonconvexity issues, Wang et al.[27]proposed a novel robust loss function to minimize misclassification cost, which is robust to the different types of noise environments, such as label noise and feature noise. Lin et al.[28] utilized the focal loss to redefine the loss function based on DNN (deep neural networks) rather than the cost matrix. Focal loss is designed to focus learning on hard examples to essentially handle the class imbalance issue in a simple and effective way. In addition, the problem of both imbalanced and incomplete datasets can be addressed by the effective combination of approaches, such as optimized k-NN approach[29] and fuzzy-based information decomposition method[30]

1.3. Our approach

As quoted above, class imbalance and incompleteness reserve to be two main obstacles in achieving the successful application of machine learning method for prediction. In the medical field, especially for stroke prediction, more attention shall be paid to the rate of false negatives and false positives among the outputs of model. In this paper, the approach of stroke prediction mainly contains two

steps. Firstly, different methods are used to impute the missing values according to the characteristics of data. Secondly, we propose a prediction model based on AutoHPO for class imbalance dataset to implement stroke prediction. Then, this paper develops a special metric for prediction and compares it with the other common algorithms. The contribution of this work is at least threefold:

(1) We propose a hybrid machine learning approach to predict stroke based on incomplete and imbalanced dataset. The approach consists of the data-level preprocessing and algorithm-level learning. The former can help undersample majority class to a manageable level instead of until to the balance. The latter can deal with slightly imbalanced dataset without manual hyperparameter tuning;

(2) Instance selection of automated hyperparameter optimization(AutoHPO) extracts hard-classified samples, which are the most similar with minority class. This method can improve the accuracy and stability of prediction;

(3) We use online-learning strategy to reweight each batch of the training set for addressing the problem of slightly imbalanced dataset. This strategy optimizes the parameters of model based on validation loss rather than training loss as usual, and achieves impressive performance on imbalanced dataset. The modified loss function are developed for a lower false negative rate and to guarantee the overall accuracy of prediction.

The reminder of this paper are organized as follows. In Section 2, we provide a formal problem description for stroke prediction with data-missing and imbalanced class. Section 3 describes the overall algorithm framework of the proposed method. In Section 4, experimental results and analysis are provided to show the effectiveness of the proposed approach. Finally, Section 5 gives some concluding remarks and suggest some potential directions for future research.

2. Problem description

In this study, we attempt to predict stroke based on a dataset only including physiological indicators, such as age, blood pressure and blood glucose, but

without any complex medical monitoring. This strategy will lower diagnosis cost, while also make it more difficult to predict. For stroke prediction, the problem essentially becomes a binary classification, which means the prognosis results are divided into stroke and non-stroke. The main objective of predictive modeling is to obtain an unknown function $\{f(\mathbf{x},\theta),\theta\in\Lambda\}$ based on a training dataset $D=\{\langle\mathbf{x}_i,y_i\rangle\}_{i=1}^{n}$, where θ is parameter in the function, $\mathbf{x}$ are feature vectors and $y\in\{0,1\}$ are binary labels. Let $L(y,f(\mathbf{x},\theta))$ be loss function between the response y to a given input vectors $\mathbf{x}$ and the response $f(\mathbf{x},\theta)$ provided by the learning-based methods. The risk functional is defined by the expected value of the loss function as follows:

$$R(\theta)=\sum_i L(y_i,f(\mathbf{x}_i,\theta)). \tag{1}$$

Then, the objective is to select an optimal function $f(\mathbf{x},\theta)$ among the unknown function set which can minimize the $R(\theta)$. Regarding the high-quality dateset shown as Fig.2(a), this standard approach is mature and valid. However, the medical dataset is usually incomplete (See Fig.2(b)) and imbalanced (See Fig.2(c)) which fails to implement the above-mentioned approach. To further clarify this issue [31], let $\mathbf{x}_i=[\mathbf{x}_i^{o_i};\mathbf{x}_i^{m_i}]$, where $\mathbf{x}_i^{o_i}$ and $\mathbf{x}_i^{m_i}$ respectively denote complete and missing samples. Then introduce two indicator matrices $[O_i]_{D_i^{o_i}\times D}$ and $[M_i]_{D_i^{m_i}\times D}$ to the samples, which satisfied $x_i^{o_i}=O_ix_i$ and $x_i^{m_i}=M_ix_i$, thus the $\mathbf{x}_i$ can be rewritten as equation (2).

$$\mathbf{x}_i=\begin{cases}\mathbf{x}_i^{o_i}, & D_i^{o_i}=D,\\ O_i^T\mathbf{x}_i^{o_i}+M_i^T\mathbf{x}_i^{m_i}, & 1\le D_i^{o_i}\le D,\end{cases} \tag{2}$$

where $O_i^TO_i+M_i^TM_i=I_D$, $D_i^{o_i}$ denotes the size of observation dimension. It is obvious to find that the incomplete dataset can affect the performance of prediction, since a portion of samples are not involved in the calculation. For the imbalanced dataset, let $1(\cdot)$ be the indicator function, if $\sum_i 1(y_i=1)<<\sum_i 1(y_i=0)$, the model will strive to reduce the overall expected risk regardless of $1(y_i=1)$ set. What's worse, it may directly lead to a serious mistake if

the "small" set is neglected, such as clinical misdiagnose and mechanical fault undetected. Therefore, we must take action to avoid such problems.

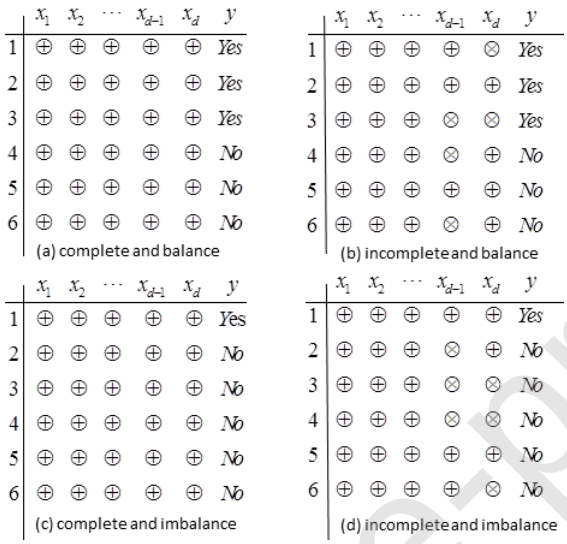


Fig. 2. 4 patterns of medical dataset

As discussed earlier, it is difficult to handle either incomplete dataset or imbalanced dataset with the standard approach, let alone the dataset with these two problems at the same time, and the dataset pattern is listed as Fig.2(d). In the Fig.2, "Yes" and "No" respectively denote stroke and non-stroke, $\otimes$ represents missing, and $\oplus$ is complete. Under this background, the main difficulty is how to predict stroke accurately and validy. Here, we put forward a solution to deal with it, and the whole procedure chart are shown as Fig.3.

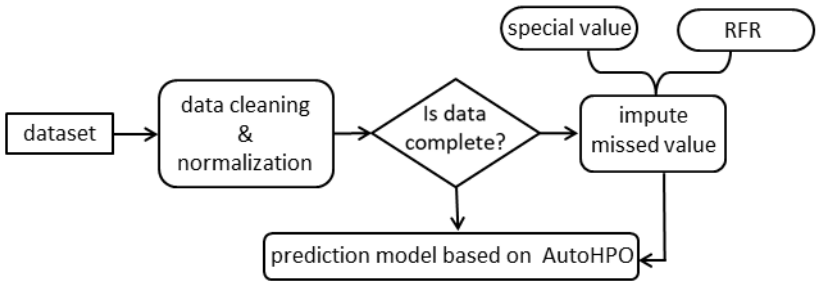


Fig. 3. Overall procedure chart

3. Material and method

3.1. Data preprocessing

The outliers and noise values among the dataset can affect the performance of classifier model. There are two popular methods to filter the noise and outliers: statistical method and non-parametric method[32]. The statistical method is simple to implement based on prior knowledge, while we utilize this method to remove the obvious noise and the outliers which are against the basic medical knowledge.

Data normalization is to eliminate dimension relations among different variables and realize data comparability through data-scaling, which helps to increases accuracy of the algorithm. In this study, we utilize Z-score to normalize the data, which can avoid the effects of extreme values through indirect centralization.

To improve computation efficiency and avoid unfavorable impact caused by irrelevant features, this paper uses filter strategy to remove the obviously irrelevant and redundant features for retaining the implicit information of the original dataset.

3.1.1. Imputation methods for missing data

Missing value imputation is a crucial work in the preprocessing with two commonly used approaches. The first is to impute special value or statistic value (e.g.mean value and median value). The second is to impute missing data

based on prediction model, such as regression and classification algorithm. If the percentage of missing values is not large, the second approach is generally superior to the first one, since the special values may seriously impact the structure of original data. However, it is extremely difficult to impute missing data by prediction in some cases if the correlation among features is weak. Therefore, we require that disparate methods should be taken based on the characteristics of the data. The correlation degree between the missing items and the other complete ones is the main factor in the decision-making process. In this study, we give preference to the prediction method for missing value imputation if the correlation among features is relatively strong, otherwise special value will be adopted. Regarding to the algorithm selection for prediction, this paper utilizes random forest regression (RFR) because it performs better than the others in accuracy and generalization. In addition, RFR does not require feature engineering, which is appropriate for the incomplete dataset.

Random Forest (RF) is a well-known ensemble algorithm in which classification is made by a set of individual decision trees. Each tree is regarded as a subset of training samples, while each node of the tree is randomly selected for several times to decrease the correlation between the trees for a lower error rate. In order to split the feature with the lowest impurity at each node, we adopt Gini criterion to select[33], which is given by:

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k), \tag{3}$$

where p_k denotes the probability of the samples belonging to class k. The label of each sample is obtained, then majority vote method is usually adopted to predict unlabeled examples. Random forest regression (RFR) is a practical application of RF, which attempts to minimize squared error loss rather than majority vote. We use the grid search to optimize the hyperparameters and then assess the prediction quality by K-fold cross-validation. It is demonstrated in the experiment that the RFR model performs better than other common algorithms (see Section 4 for more details).

3.2. *A practical prediction model based on AutoHPO for class imbalance*

As mentioned above, class imbalance can seriously impact the performance of prediction. Here, we take advantage of adaptive feature selection of AutoML[34] to undersample majority class, and redesign the AutoML framework to predict stroke. Therefore, we propose a novel classification model based on automated hyperparameter optimization(AutoHPO) to deal with our imbalanced dataset, in which the imbalance ratio is reduced by instance selection and an online learning strategy based on DNN model is designed to predict stroke. This approach consists of two phases: In the first phase, we select the optimized hyperparameter of instance selection for undersampling majority class based on its training loss. In the second phase, an online-learning step is used to reweight each batch of training set according to its validation performance. In the scenario of imbalanced dataset, this approach can ensure the results are accurate and valid without large training cost. The procedure chart of the prediction model is shown as Fig.4.

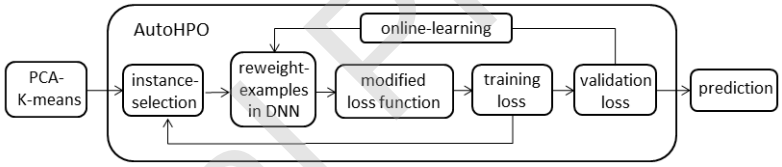


Fig. 4. Procedure chart of AutoHPO.

In Fig.4. PCA-Kmeans is the preprocess to reduce the data dimension and preliminary classification(Section 3.2.1). Instance selection is used to undersample majority class, and the outcome acts as the input of DNN(Section 3.2.2). Then, we train the hyperparameter of instance selection by the modified loss function, and the optimized instances can obtain the minimum training loss. For further improving the performance of our model, online-learning strategy is adopted to reweight examples based on the validation loss(Section 3.2.3).

3.2.1. *PCA-Kmeans for AutoHPO preprocessing*

Many medicine datasets consist of large feature space, and the high-dimensional dataset will deteriorate the performance of basic algorithms, such as K-means and K-NN[35]. Principal component analysis (PCA) is a widely multivariate statistical technique for data compression. The main function of it is to simplify the data by mapping m-dimension (original data dimensions) to n-dimension ($m > n$) based on linear transformation. We utilize PCA to compress the original dataset to a low-dimension subspace, which can be effectively clustered by K-means. K-means is one of the most effective clustering algorithms for unsupervised learning tasks, and the basic theory is to divide the sample set $D = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_s\}$ into k clusters $C = \{C_1, C_2, ..., C_k\}$, $k < s$. The minimum sum of the squared error(SSE) over all k clusters is the objective of K-means, SSE is given by:

$$J(C) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mu_k\|^2. \tag{4}$$

where μ_k is the centroid of cluster C_k. In this study, we will apply the simple and effective "Elbow" method to optimize the value k[36].

However, traditional PCA-Kmeans for undersampling is bound to randomly extract a certain amount of data from each cluster, which will lead to implicit information loss or overlap to some extent. Therefore, we use instance selection to extract samples from each cluster instead of random extraction.

3.2.2. Instance selection for undersampling

As an advanced re-sampling technique in machine learning[37], instance selection(IS) is widely adopted to filter out noisy and irrelevant data samples from the original datasets. It can also balance the data distribution and reduce the imbalance ratio (IR) in the data level for the imbalanced issue. In this study, we use k_s-NN based on instance selection to undersample the majority class of the original data. As mentioned above, we have divided sample set D into k clusters $C = \{C_1, ..., C_k\}$, and let $C'_k = \{(\mathbf{x}_i, y = 1)\}$ be the sample set of minority class and $C''_k = \{(\mathbf{x}_i, y = 0)\}$ be the majority class, $C_k = C'_k \cup C''_k$. The d_i^k denote the distance between $\{\mathbf{x}_i, \mathbf{x}_i \in C'_k\}$ and all samples $\{\mathbf{x}_i, \mathbf{x}_i \in C''_k\}$

in each k cluster of the majority class. we compute and sort the distance d_i^k, which act as the criterion of selection. The nearest neighbor is a local learning algorithm. Therefore, this strategy can reduce the calculation cost and error to some degree, and it inspired by the batch idea of deep learning where each cluster size is regards as a batch size.

In addition, we pick the nearest k_s samples as the majority class instances rather than the most distant k_s samples. This strategy can increase model's robustness and generalization, since we select these examples which are the most similar to the minority class samples and regard them as the extreme cases. Then, k_s is the hyperparameters of this algorithm, which plays a critical role in classification and reducing imbalance ratio. We adaptively optimize hyperparameter based on training loss via an instantiation of meta-learning rather than manual search. To be more precise, this strategy can significantly reduce the search scope of online learning, otherwise, we must reweight each input for the optimization results with different results of instance selection.

3.2.3. Online learning of DNN prediction model for stroke

Classifier can scarcely performs appropriately in dealing with underrepresented data and severe class distribution skews. Deep neural network (DNN), a neural network with multi-hidden layers, can express the highly nonlinear function. Under such a situation, we have designed a DNN prediction model for slightly imbalanced data by online learning strategy, which contains two parts: reweight examples and modified loss function.

Reweight examples. For the class imbalance problem, DNN is prone to training set bias, even though it has a powerful capacity for classification. Therefore, we propose a method to reweight each batch of training set based on the validation loss, and optimize each batch weight ω_j as the hyperparameters. Let $\{(\mathbf{x}_{i,j}, y_{i,j}), 1 \leq i \leq N, 1 \leq j \leq M\}$ be the training set and equally and randomly divide it into M parts, $D^v = (\mathbf{x}^v, y^v)$ is the validation set. The DNN model aims to minimize the loss for the training set, and the loss function is

defined as follows:

$$J(\theta,\omega)=\sum_{j=1}^{M}\sum_{i=1}^{N}\omega_j L_j(f(\mathbf{x}_{i,j},\theta),y_{i,j})+\psi(\theta), \tag{5}$$

$$\theta^*=\arg\min_{\theta} J(\theta,\omega), \tag{6}$$

where θ is weights and biases set of DNN, $L_j(\cdot)$ and ω_j respectively denote loss function and weights of jth batch training set, $\psi(\theta)$ is regularizer term, $f(\mathbf{x}_{i,j},\theta)$ is the prediction of the ith sample, and $y_{i,j}$ represents the label of the ith sample belongs to jth part. The imbalance ratio and noises of each batch training set are not clear, so we reweight them according to the validation loss with the aim to improve the generalization performance. For the $\{\omega_j\}_{j=1}^{M}$ optimal selection, we aim to minimize the validation loss[38]:

$$\omega^*=\arg\min_{\omega_j,\omega_j\geq 0}\frac{1}{M}\sum_{j=1}^{M}\omega_j L_j^v(f(\mathbf{x}^v,\theta^*),y^v)), \tag{7}$$

where $L^v(\cdot)$ is validation loss expectation.

Modification of loss function. As a typical loss function in the neural network, cross-entropy loss function represents the discrepancy between the actual output and desired output. $p(t)$ and $q(t)$ denote two probability distributions, and the basic cross-entropy function is listed as equation(8)[39]:

$$H(p,q)=-\sum_{t}p(t)\log q(t). \tag{8}$$

The smaller the cross-entropy value is, the closer the probability distribution of $p(t)$ and $q(t)$ will be. According to equation(8), the cross-entropy loss function for binary classification can be rewritten below :

$$L^1(\theta)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{M}y_{i,j}\log P(y_{i,j}=1\mid\mathbf{x}_i,\theta)+\psi(\theta), \tag{9}$$

$$L^2(\theta)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{M}y_{i,j}\log P(y_{i,j}=0\mid\mathbf{x}_i,\theta)+\psi(\theta), \tag{10}$$

$$L(\theta)=L^1(\theta)+L^2(\theta), \tag{11}$$

where $L^1(\theta)$ and $L^2(\theta)$ respectively denote the loss of different classes. For class imbalance issue, classifier is aimed at validly recognizing the minority samples. For this reason, it is feasible to augment the minority misclassification loss to avoid minority class being neglected in the prediction model, which will impact the accuracy of majority classification to some degree. The loss function is finally adjusted as follows:

$$L(\theta) = \frac{\lambda}{\lambda+1} L^1(\theta) + \frac{1}{\lambda+1} L^2(\theta), \tag{12}$$

where λ is penalty coefficient of the minority class misclassification, and the value of λ is dynamically adjusted based on the whole training set imbalance ratio.

Optimization strategy. For the training set, the variant SGD(stochastic gradient descent) is used to train the parameters θ of models. Unfortunately, for the validation set and hyperparameter ω, we cannot directly calculate each single loop to optimize it, which is limited to the expensive cost and poor efficiency. So we utilize the online-learning strategy to select optimal ω based on the performance of validation loss. Let ϵ_j be the perturbing term of each batch weight at the training step t, and $\theta_{t+1}(\epsilon) = \theta_t - \alpha \nabla \epsilon_j L^v(\theta_t)$, where $L^v(\cdot)$ denotes validation loss function and α is the step size. Let ϵ^* denotes the locally minimal loss of validation at each step t, which is defined as:

$$\omega_{j,t} = \max\left\{(-\eta \frac{\partial}{\partial \epsilon_{j,t}} L^v(\theta_{t+1}(\epsilon))), 0\right\}, \tag{13}$$

where η is the descent step size on ϵ, and $\sum_{j=1}^{M} \omega_j = 1, \omega_j \geq 0$. During training, if the gradient direction of the training batch is similar to the validation batch, the weight of this batch should be up. In this study, we can calculate the gradient of $L^v(\cdot)$ by automatic differentiation, which is a well-known and easily practical technique. The training time of online learning has required an approximately tripling of the traditional method, so we utilize optimization hyperparameter of instance selection to reduce computational complexity(see Section 3.2.2). Ren et al.[38] have proved the convergence of the reweighted algorithm.

In addition, we attempt to reduce the number of parameters on the basis

of ensuring the validity and accuracy of model. Therefore, the basic DNN model contains 3 hidden layers, and activation functions are rectified linear units (Relu) in this study. The numerical experiment results can clarify the suitability of 3 hidden layers. For the characteristics of stroke prediction, we take the specific assessment metrics for the model performance, which is attaches more importance to the false negative ratio, and the details will be described in experiment part.

Algorithm 1 Prediction model based on AutoHPO.

Input: $\theta_0, D^v, C'_k, C''_k, N, M$;
Output: k_s, θ_T;

1: **for** k_s in $\{k_1, k_2, ..., k_c\}$ **do**
2: Calculate and sort $d_i^k, i = 1, 2, ..., k$;
3: $D_n \leftarrow$ Select the k_s nearest neighbours as new majority class ;
4: $\lambda \leftarrow$ imbalance ratio of D_{new}
5: $(\mathbf{x}_i, y_i) \leftarrow$ each bath (D_n, N) ;
6: $L(\theta_t) \leftarrow \frac{\lambda}{\lambda+1} L^1(\theta_t) + \frac{1}{\lambda} L^2(\theta_t)$;
7: **for** $t = 1$ to epochs **do**
8: $\nabla\theta_t \leftarrow \frac{\partial}{\partial \theta_{t,i}} \frac{1}{N} \sum_{i=1}^{N} L_i(\theta_t)$;
9: $\widehat{\theta}_{t+1} \leftarrow \theta_t - \alpha \nabla \theta_t$;
10: **end for**
11: **end for**
12: $\widehat{k}_s, \widehat{D}_n \leftarrow \underset{(k_s, D_n)}{\text{argmin}} L(\theta_t)$
13: Initialization: $\epsilon_j \leftarrow 0$;
14: **for** $t = t + 1, .., T - 1$ **do**
15: $(\mathbf{x}_i^v, y_i^v) \leftarrow$ each bath $(\widehat{D}_n^v, M)$;
16: $L^v(\theta) \leftarrow$ forwad $(\mathbf{x}_i^v, y_i^v, \widehat{\theta}_{t+1})$;
17: $\nabla\epsilon_j \leftarrow -\eta \frac{\partial}{\partial \epsilon_{j,t}} L^v(\theta_{t+1}(\epsilon))$;
18: $\omega_j \leftarrow \max\left\{(-\nabla\epsilon_j, 0)\right\}$
19: $\widehat{L}^v \leftarrow \sum_{j=1}^{M} \omega_j L_j^v(\theta_{t+1}(\widehat{\epsilon}))$;
20: $\nabla\theta_{T-1} \leftarrow \frac{\partial}{\partial \theta_{T-1,j}} \frac{1}{M} \sum_{j=1}^{M} \widehat{L}_j^v$;
21: $\widehat{\theta}_T \leftarrow \theta_T - \alpha \nabla \theta_{T-1}$
22: **end for**
23: **return** ω_j, θ_T

The training process of our approach is presented in Algorithm.1. The model has two stages: 1) instance selection reduces the imbalance ratio of dataset by undersampling majority class(line 1-12). 2) online-learning strategy is used to deal with slightly imbalanced dataset(line 13-23). In the first stage, the inputs of instance selection C'_k and C''_k are obtained by the data-level preprocessing, and

the search scope of k_s is the set $\{k_1, k_2, ...k_c\}$. We pick the nearest k_s samples to get an updated dataset, and the objective k_s value denotes the minimum training loss which is optimized by the backpropagation training algorithm(line 7-10). In the second stage, the initial validation loss is calculated by the forward propagation algorithm(line 16). Then, we add perturbing ϵ_j for each batch in DNN to rectify the weights ω_j and ensure the weights are non-negative(line 18-19). For each training iteration, the backpropagation algorithm is also used to reweight each batch according to the performance of validation loss(line 20-22).

4. Experiment and result

4.1. Dataset

In this study, the original dataset of stroke is collected from HealthData.gov, which is also utilized as the benchmark dataset in a Kaggle competition[2]with details listed as Table 1. The dataset is a typical class imbalanced type and contains 11 features, where 783 occurrences of stroke were included in a total of 43400 recorded samples, only accounting for 1.18% of the whole. Furthermore, the dataset is incomplete, and 30% smoking status items and 3% body mass index (BMI) items are missing. For such data missing issue, different methods will be adopted in the preprocessing.

4.2. Experimental setup

4.2.1. Prediction with DNN based on AutoHPO model

The first step of AutoHPO is to reduce the data dimension by PCA, and after the principal components are extracted, the selection of k value is crucial to K-means performance. Considering the large size of this stroke dataset, we take advantage of "Elbow" method to evaluate the k selection. This strategy can reduce the calculation cost and complexity of the post-processing algorithm, and it also can avoid overlapping of the sampled dataset.

[2]https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

Features	values	Features	values
Patient ID	1-43400	Hypertension(hyp)	Yes/No
Gender(gen)	Male/Female	Married(mar)	Yes/No
Residence type	Urban/Rural	Age	0.08-82
Avg-glucose(glu)	55-291	Heart disease(htd)	Yes/No
Work type(work)	Private/Employed	BMI	10.1-97.6
Smoking status	Smoked/Formerly/Never		

Table 1: Dataset description

Instance selection is used to reduce the imbalance ratio by undersampling of the majority class according to the result of k_s-NN algorithm. For the selection of hyperparameter k_s values, we dynamically select it by the training loss of the prediction, and hyperparameter k_s will be sampled from {3,5,7,9,11}. There are at least two reasons for the limitation of the search scope. Firstly, k_s is usually odd number in order to reduce classification error rate. Secondly, the stroke dataset is obviously imbalanced, we just reduce the imbalance ratio to a manageable size instead of re-sampling until to the class balance which will seriously changes the original distribution. In addition, each batch weight is equal in this part, and the optimal selection of k_s should ensure that the false negative rate of prediction outcome is the minimum and the overall accuracy of the outcome is more than 70%.

Online learning strategy of reweight examples is implemented based on a DNN model with 3 hidden layers, whose activation functions are the rectified linear units (Relu) and the output layer is the 2-dimension softmax function. The initial weight-examples of each batch are equal with the sum of 1. We will use automatic differentiation to optimize and reweight the examples. In view of the serious consequences caused by medical misdiagnosis, it needs to modify loss function, the weight coefficient λ of minority classes is set as the imbalance ratio after instance selection.

The objective functions of this model are adjusted to sensitivity and speci-

ficity as the equation (17) and (18). To improve the accuracy of prediction outputs, 10-fold cross-validation is necessary for our model training to avoid overfitting problem. Without loss of generality, it is essential to repeat the experiments of the overall prediction model for 3 times. Then, we take the mean value of these outputs as the final prediction output.

As for stroke prediction, the common assessment metrics can provide reference for the overall accuracy assessment of a prediction model, which is given by equation (15). Then we propose the specific assessment metrics and regard them as the evaluation functions of our prediction model.

4.2.2. Assessment metrics for class imbalance

To resolve the problem of class imbalance, the four metrics based on prediction results will form a confusion matrix, and they are respectively true positive (TP), false negative (FN), false positive (FP) and true negative (TN)[40]. The assessment metrics are adjusted to sensitivity, specificity and G-Mean, which are defined respectively as equation (16)-(18). To be specific, G-Mean is used to assess the balance degree of the algorithm for class imbalanced data, and the larger the value is, the better it will be. In addition, receiver operating characteristic (ROC) curve, an effective visual tool will be adopted to visualize the assessment metrics.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{14}$$

$$sensitivity = \frac{TP}{TP + FN}, \tag{15}$$

$$sepecificity = \frac{TN}{TN + FP}, \tag{16}$$

$$G\text{-}Mean = \sqrt{sensitivity \times sepecificity}. \tag{17}$$

4.2.3. Assessment metrics of stroke prediction

The assessment metrics above in the previous section all take *accuracy, sensitivity* and *G-Mean* to objectively evaluate the classifier to some degree. However, they are inapplicable in this paper. The medical diagnosis should give

priority to false negative rate (FN_{rate}) and false positive rate (FP_{rate}), especially for stroke prediction. Against the background, the paper adopts false negative rate and false positive rate as the main assessment metrics for stroke prediction in the premise of accuracy. The evaluation functions can be defined as below[41]:

$$FN_{rate} = \frac{FN}{TN + FN}, \tag{18}$$

$$FP_{rate} = \frac{FP}{TP + FP}. \tag{19}$$

4.3. Computational results and analysis

4.3.1. Results of data preprocessing

In this subsection, we will use the stroke dataset to verify the prediction method for missing values in Section 3. This dataset contains some obvious outliers and noises, such as age and BMI items. According to the methods and standards from MONICA [3][42], the minimum age of stroke-monitoring should be 25. Unfortunately, some samples younger than 25 are also included in the dataset, and even the infant with only 0.08 ages. The reference values of BMI normally range from 10% to 50%, but some samples with BMI value between 60% and 97.6% are also covered in the collected dataset. Hence such data (less than 25 in ages and more than 60% in BMI) should be removed as outliers and noises. Feature selection is the process to remove the irrelevant and redundant features through filter methods. The distributions of the residential-type items are almost the same with each other. Patient ID item is an obvious redundancy feature. Such features should be directly removed.

4.3.2. Imputing of missing values

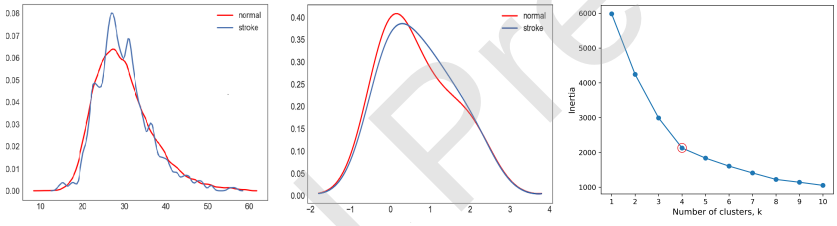
Regarding missing value imputation, different strategies are utilized based on the characteristics of various features. According to the correlation analysis, it has been found that the correlation between smoking status item and the

[3]Multinational Monitoring of Trends and Determinants in Cardiovascular Diseases

	gen	age	hyp	htd	mar	work	glu
bmi	0.023	**0.11**	**0.13**	0.025	**0.15**	-0.036	**0.18**
smoke	0.091	0.061	0.008	0.063	0.07	-0.023	0.026

Table 2: The correlation coefficient of features

other features is weak, while the correlation between BMI and others is relatively strong, as shown in Table 2. For the stroke dataset with obvious class imbalance, the simple correlation analysis fails to describe the correlation between the features and the label item. The necessary of the prediction method should be conducted by comparing the distribution between feature and target. In Fig.5 (a) and (b), it can be seen that for different target values, BMI has relatively obvious distinction, while smoking status item does not.



(a) BMI (b) Smoking status (c) Selection of K

Fig. 5

Table 3 also illustrates that the commonly used prediction methods are not desirable for the smoking status item, such as k-NN, logistic regression (LR), RF and decision tree (DecT), thus special value imputation is adopted for smoking status item, and the percentage of missing values of the smoking status is too high. Thus special value imputation is adopted for it. For BMI item, RFR acts as the prediction and imputation method, and hyperparameters in the RFR model are optimized through the grid search. The prediction results will be compared with those classical machine learning methods, including SVM,

Algorithm	k-NN	LR	RF	DecT
Accuracy	0.496	0.551	0.531	0.553

Table 3: Prediction accuracy of smoking status

Algorithm	SVM	RFR	BayRid	GBM
RMSE	0.641	0.724	0.719	0.648
R-squared	0.301	0.876	0.184	0.326

Table 4: Prediction evaluation of BMI

Bayesian-Ridge (BayRid) and Gradient-Boosting (GBM). The main evaluation metrics of the prediction are R-squared and RMSE (Root Mean Square Error). R-squared value refers to the proportion of regression sum of squares in the total sum of squares in multiple regression, which is the statistical result of fitting degree. The closer to 1 the R-squared value is, the better will be. RMSE represents the error of the regression. This evaluation metric is sensitive to the outlier and the result of it is non-robust. The prediction results in Table 4 indicate that RMSE value from RFR is slightly higher than others, but the R-squared value obtained from RFR is much better than other algorithms.

4.3.3. Results of DNN based on AutoHPO model

The numerical experiment is designed to verify the hybrid machine learning approach in this paper. PCA is used to reduce the original data to 5 dimensions with 89% information left. The k value of K-means is 4, which is regarded as the "elbow " point as shown in Fig.5(c). For verification of the false negative rate of prediction with different instance selections, the outputs have been collected from the model of our approach, DNN(without reweight-examples), Bagging(Bag), RF, Adaboost(Ada) and XGBoost(XGB) for comparison, as shown in Fig.6. Further comparison of ROC curves is provided to assess the overall accuracy of models, and the false positive rate has been reduced with the decrease of imbalance ratio.

To guarantee the overall accuracy of other traditional approaches, such as Ada and XGB, the value k_s of instance selection is selected as 5 to undersample majority class, and the output of undersampling acts as the predicted dataset of these approaches, and the prediction of our model are listed in Table 5. According to the numerical experiment results, false negative rate from our approach is only 19.1%, while accuracy reaches 71.6%. It can be concluded that the method we propose is more valid and accurate. On the one hand, the model with reweight-examples is more stable and robust than the model without it, especially suitable for the extreme imbalance dataset. On the other hand, the false negative rate from our model is distinctly much lower than those from other algorithms, and even though the false positive rate is slightly higher than others, it is still in the acceptable scope. What's more, the overall accuracy is reliable which only reduces by 1.7% in comparison with the mean accuracy of other well-known methods, and ROC performance also support this point, as shown in Fig.7 (a). The primary cause of above results lies in that traditional classifier prediction is designed for the balanced and unbiased samples, and majority classes are given much more attention in the imbalanced cases. Under such conditions, it will result in a lower false positive rate and higher false negative rate than the actual situation if we pursue the classification performance of majority classes. The G-mean of our approach is 46.9% which is the largest one among the counterparts of different methods, we observe that our model is more balanced.

	FN_{rate} (%)	FP_{rate} (%)	Accuracy(%)	Specificity(%)	Sensitivity (%)	G-mean (%)
Ours	**19.1** ±1.7	33.1 ± 0.5	71.6 ± 1.2	**32.6** ± 0.5	67.4 ± 0.5	**46.9** ± 0.5
DNN	24.8 ± 2.2	36.1 ± 1.1	65.1 ± 1.1	32.3 ± 0.5	67.7 ± 0.5	46.7 ± 0.5
Bag	70.0 ± 1.4	11.2 ± 0.2	73.8 ± 0.2	12.4 ± 0.5	87.6 ± 0.5	32.9 ± 0.5
RF	71.4 ± 2.9	10.7 ± 1.4	72.8 ± 0.8	10.7 ± 0.5	89.3 ± 0.5	30.9 ± 0.5
XGB	69.0 ± 0.7	9.5 ± 0.2	74.1 ± 0.4	11.3 ± 0.2	88.7 ± 0.2	31.7 ± 0.2
Ada	72.0 ± 0.7	10.5 ± 0.2	72.6 ± 0.4	10.5 ± 0.2	89.5 ± 0.2	30.7 ± 0.2
Avg	70.6 ± 1.6	10.5 ± 0.5	73.3 ± 0.5	11.2 ± 0.4	88.8 ± 0.4	31.6 ± 0.4

Table 5: Assessment values obtained for different approaches

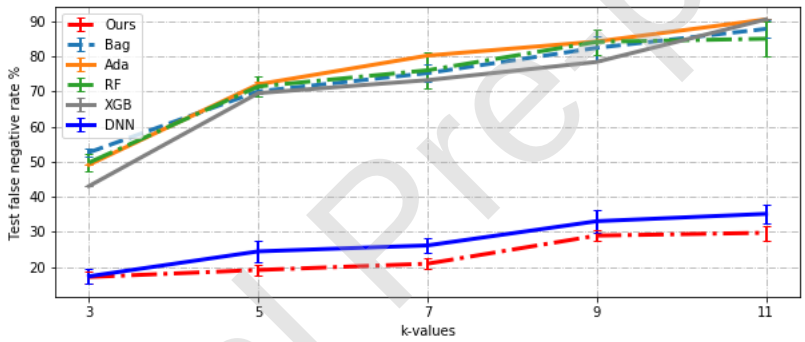


Fig. 6. The false negative rate with different imbalance ratios

In addition, statistical significance is needed to assess the model performances. The null hypothesis is that there is no difference of the FN_{rate} between our approach and the other well-known ones, whereas the alternative hypothesis is that there is difference between them. Here, the paired t-tests is used to evaluate whether the difference is significant enough to reject null hypothesis. The calculation results of significance testing show that our approach effect is statistically significant (as listed in Table.6). [4]

[4] According to the traditional significance criterion: $p < 0.001$ extremely significant, and $p > 0.05$ not significant.

	Mean	t-value	p-value	statistical significance
RF	0.61	-9.54	$4.36 * 10^{-16}$	Extremely significance
Bag	0.33	-11.94	$1.39 * 10^{-21}$	Extremely significance
Ada	0.50	-5.07	$1.63 * 10^{-6}$	Extremely significance
XGB	0.62	-17.45	$1.77 * 10^{-33}$	Extremely significance

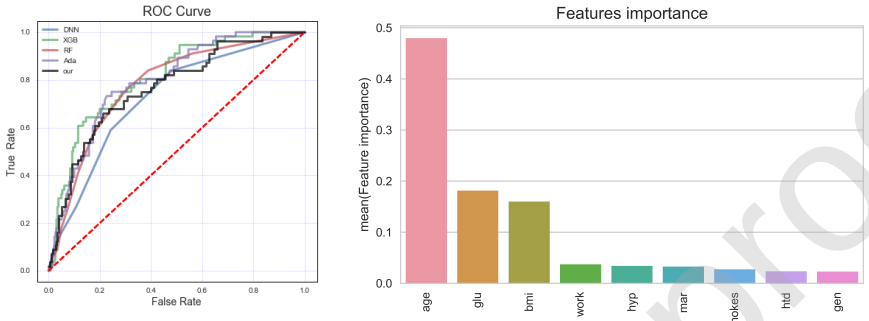
Table 6: Results of the significance testing by paired t-test

4.3.4. Result analysis of stroke prediction

The data of this study are collected through non-invasive detection, and the target is to predict the probability of stroke-suffering and the necessity to prevent in advance at the lowest cost. Therefore, it should reduce the false negative rate at first. On the other hand, if the false negative rate is too high, it will seriously interfere with clinical diagnosis, and even fail to take preventive measures because of misdiagnosis, which will lead to irreversible damage to patients. What's more, the psychological factor can also impact the stroke-suffering rate. High false positive rate can impose a heavy psychological burden on the sick, which is unfavorable to their health. According to the analysis above, the prediction in our approach is obviously superior to other well-known ones.

In addition, we obtain the importance of various features by the traditional method, since the DNN can not perform well for their analysis. Here, the importance of degrees in Fig.7(b) are the mean values of XGB and RF analysis results. In order to guarantee the accuracy of feature importance, the input data for these methods are the balance date from instance selection in this paper. Fig.5(b) demonstrates that major factors for stroke are age, blood glucose, BMI, hypertension and heart disease, which are basically consistent with analysis result of traditional medicine. It is noted that factors such as blood glucose, BMI and hypertension can be adjusted through the healthy lifestyle, and heart diseases can also be controlled by drugs. Considering that the higher false

negative rate from the traditional method is undesirable, the feature importance from the traditional method can only serve as the clinical reference instead of clinical evidence. Feature importance analysis by DNN models deserves further research for better prediction in the future.



(a) (b)

Fig. 7

5. Conclusion

To improve medical prediction based on physiological indicators of potential stroke patients, this paper has proposed a hybrid method which combines missing value imputation and a AutoHPO-based DNN prediction model. The outcome of our model is that the false negative rate is only 19.1% and overall accuracy is 71.6%. In comparison with the mean results of other commonly used methods, the false negative rate decreass by 51.5% and the overall error increases by 1.7%. Such changes mean that our approach can substantially reduce false negative rate without a large cost of overall accuracy. Therefore, the hybrid machine learning approach in this study is effective and credible for stroke prediction. In addition, this approach can dynamically optimize the hyperparameter without manual selection, and it also considers the correlation of multi-factors, which is more advanced than single-factor analysis commonly used in traditional medicine.

As for the stroke prediction, further studies can be conducted on the feature importance analysis based on deep neural network, such as sensitivity analysis and ℓ_1 regularization, which help specify physiological indicators for targeted controlling. With these efforts, further improvements will be achieved for more valid and mature stroke prediction.

References

[1] V. L. Feigin, B. Norrving, G. A. Mensah, Global burden of stroke, Circulation research 120 (3) (2017) 439–448.

[2] M. Naghavi, K. M. Abajobir, F. Abd-Allah, A. Abera, et al., Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016, The Lancet 390 (10100) (2017) 1151–1210.

[3] A. Algra, M. J. Wermer, Stroke in 2016: Stroke is treatable, but prevention is the key, Nature Reviews Neurology 13 (2) (2017) 78.

[4] M. J. O'Donnell, S. L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang, P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay, et al., Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (interstroke): a case-control study, The Lancet 388 (10046) (2016) 761–775.

[5] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, L. Hua, Data mining in healthcare and biomedicine: a survey of the literature, Journal of medical systems 36 (4) (2012) 2431–2448.

[6] A. N. Richter, T. M. Khoshgoftaar, A review of statistical and machine learning methods for modeling cancer risk using structured clinical data, Artificial intelligence in medicine 90 (2018) 1–14.

[7] C. R. Pereira, D. R. Pereira, S. A. Weber, C. Hook, V. H. C. de Albuquerque, J. P. Papa, A survey on computer-assisted parkinson's disease diagnosis, Artificial intelligence in medicine 95 (2018) 48–63.

[8] A. Kaya, Cascaded classifiers and stacking methods for classification of pulmonary nodule characteristics, Computer methods and programs in biomedicine 166 (2018) 77–89.

[9] O. R. Shishvan, D.-S. Zois, T. Soyata, Machine intelligence in healthcare and medical cyber physical systems: A survey, IEEE Access 6 (2018) 46419–46494.

[10] C. Colak, E. Karaman, M. G. Turtay, Application of knowledge discovery process on the prediction of stroke, Computer methods and programs in biomedicine 119 (3) (2015) 181–185.

[11] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, H. Lee, An integrated machine learning approach to stroke prediction, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 183–192.

[12] N. Kasabov, V. Feigin, Z.-G. Hou, Y. Chen, L. Liang, R. Krishnamurthi, M. Othman, P. Parmar, Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke, Neurocomputing 134 (2014) 269–279.

[13] A. K. Arslan, C. Colak, M. E. Sarihan, Different medical data mining approaches based prediction of ischemic stroke, Computer methods and programs in biomedicine 130 (2016) 87–92.

[14] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artificial intelligence in medicine 50 (2) (2010) 105–115.

[15] T.-P. Hong, C.-W. Wu, Mining rules from an incomplete dataset with a high missing rate, Expert Systems with Applications 38 (4) (2011) 3931–3936.

[16] X. Zhang, S. Song, C. Wu, Robust bayesian classification with incomplete data, Cognitive Computation 5 (2) (2013) 170–187.

[17] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Systems with Applications 73 (2017) 220–239.

[18] N. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets. sigkdd explor newsl 6: 1–6 (2004).

[19] G. E. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD explorations newsletter 6 (1) (2004) 20–29.

[20] Y.-M. Chyi, Classification analysis techniques for skewed class distribution problems, Department of Information Management, National Sun Yat-Sen University.

[21] 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.

[22] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, Information Sciences 409 (2017) 17–26.

[23] Y. Wang, L. Yang, Q. Ren, A robust classification framework with mixture correntropy, Information Sciences 491 (2019) 306–318.

[24] L. Yang, H. Dong, Robust support vector machine with generalized quantile loss for classification and regression, Applied Soft Computing 81 (2019) 105483.

[25] A. Ghazikhani, R. Monsefi, H. S. Yazdi, Online cost-sensitive neural network classifiers for non-stationary and imbalanced data streams, Neural Computing and Applications 23 (5) (2013) 1283–1295.

[26] H. Yu, C. Mu, C. Sun, W. Yang, X. Yang, X. Zuo, Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data, Knowledge-Based Systems 76 (2015) 67–78.

[27] Y. Wang, L. Yang, A robust loss function for classification with imbalanced datasets, Neurocomputing 331 (2019) 40–49.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE transactions on pattern analysis and machine intelligence.

[29] E. C. Ozan, E. Riabchenko, S. Kiranyaz, M. Gabbouj, An optimized k-nn approach for classification on imbalanced datasets with missing data, in: International Symposium on Intelligent Data Analysis, Springer, 2016, pp. 387–392.

[30] S. Liu, J. Zhang, Y. Xiang, W. Zhou, Fuzzy-based information decomposition for incomplete and imbalanced data learning, IEEE Transactions on Fuzzy Systems 25 (6) (2017) 1476–1490.

[31] C. A. Leke, T. Marwala, Introduction to missing data estimation, in: Deep Learning and Missing Data in Engineering Systems, Springer, 2019, pp. 1–20.

[32] I. Ben-Gal, Outlier detection, in: Data mining and knowledge discovery handbook, Springer, 2005, pp. 131–146.

[33] A. M. Prasad, L. R. Iverson, A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction, Ecosystems 9 (2) (2006) 181–199.

[34] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, in: Advances in neural information processing systems, 2015, pp. 2962–2970.

[35] C. Ding, X. He, K-means clustering via principal component analysis, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 29.

[36] T. M. Kodinariya, P. R. Makwana, Review on determining number of cluster in k-means clustering, International Journal 1 (6) (2013) 90–95.

[37] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, G.-T. Yao, Under-sampling class imbalanced datasets by combining clustering analysis and instance selection, Information Sciences 477 (2019) 47–54.

[38] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: International Conference on Machine Learning, 2018, pp. 4331–4340.

[39] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.

[40] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (4) (2012) 463–484.

[41] P. Branco, L. Torgo, R. P. Ribeiro, A survey of predictive modeling on imbalanced domains, ACM Computing Surveys (CSUR) 49 (2) (2016) 31.

[42] M. Liu, B. Wu, W.-Z. Wang, L.-M. Lee, S.-H. Zhang, L.-Z. Kong, Stroke in china: epidemiology, prevention, and management strategies, The Lancet Neurology 6 (5) (2007) 456–464.