Intelligence-Based Medicine 10 (2024) 100167

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Machine-learning-enabled prognostic models for sepsis

Chunyan Li [a], Lu Wang [b], Kexun Li [b,c], Hongfei Deng [b], Yu Wang [b], Li Chang [b], Ping Zhou [b], Jun Zeng [b,c], Mingwei Sun [b,c], Hua Jiang [b,c], Qi Wang [a,*]

[a] Department of Mathematics, University of South Carolina, Columbia, SC, 29208, USA
[b] Institute for Emergency and Disaster Medicine, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, 610072, China
[c] Sichuan Provincial Research Center for Emergency Medicine and Critical Illness. Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, 610072, China

ARTICLE INFO

Keywords:
Sepsis
Classifiers
Class-imbalance data
Dimension reduction
Sampling strategies
Small dataset

ABSTRACT

Background and Objectives: Sepsis is a leading cause of mortality in intensive care units (ICUs). The development of a robust prognostic model utilizing patients' clinical data could significantly enhance clinicians' ability to make informed treatment decisions, potentially improving outcomes for septic patients. This study aims to create a novel machine-learning framework for constructing prognostic tools capable of predicting patient survival or mortality outcome.

Methods: A novel dataset is created using concatenated triples of static data, temporal data, and clinical outcomes to expand data size. This structured input trains five machine learning classifiers (KNN, Logistic Regression, SVM, RF, and XGBoost) with advanced feature engineering. Models are evaluated on an independent cohort using AUROC and a new metric, γ, which incorporates the F1 score, to assess discriminative power and generalizability.

Results: We developed five prognostic models using the concatenated triple dataset with 10 dynamic features from patient medical records. Our analysis shows that the Extreme Gradient Boosting (XGBoost) model (AUROC = 0.777, F1 score = 0.694) and the Random Forest (RF) model (AUROC = 0.769, F1 score = 0.647), when paired with an ensemble under-sampling strategy, outperform other models. The RF model improves AUROC by 6.66% and reduces overfitting by 54.96%, while the XGBoost model shows a 0.52% increase in AUROC and a 77.72% reduction in overfitting. These results highlight our framework's ability to enhance predictive accuracy and generalizability, particularly in sepsis prognosis.

Conclusion: This study presents a novel modeling framework for predicting treatment outcomes in septic patients, designed for small, imbalanced, and high-dimensional datasets. By using temporal feature encoding, advanced sampling, and dimension reduction techniques, our approach enhances standard classifier performance. The resulting models show improved accuracy with limited data, offering valuable prognostic tools for sepsis management. This framework demonstrates the potential of machine learning in small medical datasets.

1. Introduction

Sepsis is one of the biggest threats of death for critically ill patients in the intensive care unit (ICU) [1]. Traditionally, the prognosis of sepsis relies on the personal judgment of clinicians based on the longitudinal monitoring of a set of indicators. In 2016, the latest guideline (Sepsis 3.0) updated the definition of sepsis and defined it as a complex disease involving severe inflammatory responses and multiple organ failures; and it also pointed out that it would be difficult to obtain a positive prognosis via traditional statistical methods based on the patient's cross-sectional data at admission or at a few selected cross-sectional moments during the treatment [2]. Hence, the patient's longitudinal data, with hidden temporal features, should be exploited to provide "insightful dynamical" information for making more accurate, data-driven, intelligent, and reliable prognoses for septic patients.

Artificial intelligence (AI) fueled by the latest development in machine learning (ML) has shown great promise in medicine and health care in general. Some ML tools have been developed based on cross-sectional data in the study of sepsis so far. However, meta-analyses

* Corresponding author.
E-mail addresses: jianghua@uestc.edu.cn (H. Jiang), qwang@math.sc.edu (Q. Wang).

https://doi.org/10.1016/j.ibmed.2024.100167
Received 22 April 2023; Received in revised form 1 September 2024; Accepted 9 October 2024
Available online 28 October 2024

showed that most of these studies aimed at early diagnosis while some of them made less reliable clinical prognoses for sepsis, unfortunately [3]. Given that data acquired from septic patients are often represented by a large number of static and dynamic indicators, describing the long course of the disease with a dynamically changing complex context becomes necessary. This often makes the dimension of the variable space relative to the number of data points very high. It would require a virtually impossibly large sample size to obtain positive results with the existing prognostic models.

At the ICU, the electric data collection system is not well-tuned, nor data collection protocol unified such that the quality of the collected data is often difficult to control when used in AI/ML tools directly [4]. The amount of effective data per dimension in the data space must be increased to obtain clinically usable prognostic tools via existing methodologies. In reality, it is clinically prohibitive to collect a large set of data for a septic patient at the ICU anywhere. So, one must innovate in AI/ML methodologies to account for the clinically available, invaluable small datasets, where the ratio of data size to the indicator's dimension is relatively small, to build the prognostic tool. This motivates the current study.

Small dataset problems are often compounded by challenges such as class imbalance and high dimensionality, which have long posed difficulties for the machine learning community. The first paper on machine learning with small datasets was published in 1995 [5]. However, other studies on this topic did not emerge until 2016, when researchers began exploring the applications of machine learning across various fields [6]. Today, while there have been a few studies addressing small dataset problems in the medical field [7], most have focused on small-sized, high-dimensional, and class-balanced datasets [8].

In real-world scenarios, datasets often exhibit class imbalance, where certain classes are significantly underrepresented compared to the others. This imbalance leads to the "class imbalance" problem, also known as the "curse of imbalanced datasets", which refers to the difficulty of learning from classes with a limited number of samples [9]. Imbalanced data can significantly compromise the learning process, as most standard machine learning algorithms assume a balanced class distribution or equal misclassification costs [10]. The class imbalance problem is frequently encountered in medical diagnosis and has been recognized as one of the top 10 challenges in data mining and pattern recognition [11,12].

In this study, we address the small dataset problem in the context of sepsis, focusing on high-dimensional and imbalanced datasets. Our goal is to establish a modeling framework capable of handling such datasets, enabling the development of a reliable AI-enabled prognostic tool for predicting patient outcomes based on clinically collected longitudinal data from septic patients.

2. Materials and methods

2.1. Data acquisition

The data were collected from qualified septic patients admitted to the ICU over a period of four years. The inclusion criteria of the patients are defined as follows:

1. adult patients (age ≥ 18),
2. patients with the APACHE II score ≥ 10 on admission.

The exclusion criteria are given as follows:

1. underage patients (age < 18 years old),
2. patients with extracorporeal membrane oxygenation or renal replacement therapy during the hospitalization,
3. patients were pregnant or breastfeeding,
4. patients were participating in other clinical trials.

The dataset incorporates relevant indicators and features of sepsis for each patient, encompassing both static and dynamic variables. Static variables include basic patient biometrics, while dynamic variables consist of physiological and biochemical indicators collected daily from various clinical information systems, such as Electronic Medical Records (EMR), Hospital Information Systems (HIS), and Laboratory Information Systems (LIS).

Based on specific inclusion and exclusion criteria, the study incorporated 174 septic patients admitted between January 2018 and December 2021 (as illustrated in Fig. 1). The cohort was divided into two groups based on 28-day clinical outcomes: 84 patients in the mortality group and 90 in the survival group. For external validation purposes, data from an additional 21 patients, admitted between January 2021 and May 2022, were collected as a new dataset for testing.

This study has been reviewed by the Medical Ethics Committee of Sichuan Provincial People's Hospital (No. 266 in 2021) and registered in the China Clinical Trial Registration Center (Clinical Registration No.: ChiCTR2200056316). Since it was an observational study on historical data and would not interfere with the treatment plan of the patients, the ethics committee agreed to waive the informed consent.

2.2. Data preprocessing

The collected time series samples in the dataset with more than 30% missing entries are deleted to ensure the authenticity of the data. If a patient has multiple admission records, only the first is used. In the remaining, accepted dataset, the missing data of static variables are filled by mean values and those of dynamical variables are interpolated using cubic splines in time.

Notice that most data do not follow the normal distribution. Hence, we conduct the correlation analysis for the data using the Spearman method to select 17 dynamic indicators/features and 9 static ones (see A.1) in the dataset to train the models, in which a strong correlation between the variables is assumed if the correlation coefficient ≥ 0.7 over all 14 days. To prepare the data for classifiers, categorical features are represented by one-hot encoding, non-categorical discrete static features are normalized by the maximum and minimum of the dataset in 14 days, and dynamic features are standardized by the mean and standard deviation of the data in 14 days.

2.3. Performance assessment

The area under the receiver operating characteristic curve (AUROC) and the F1 score are proper metrics for assessing the models with respect to imbalanced data while the F1 score is more sensitive to gauge overfitting [13]. Hence, we use AUROC and the F1 score to assess the accuracy and the generalization ability of the classifiers, respectively. We use a metric, γ, defined below, to evaluate improvement in overfitting based on F1 scores. The first term in γ assesses how much the degree of overfitting is reduced compared to the original data without using feature engineering. The second and third term measure how much the F1 score is improved compared to the original case in internal and external validations, respectively.

$$\gamma = \left(\frac{\delta F1^{OO} - \delta F1^{ij}}{\delta F1^{OO}} + \frac{F1_{in}^{ij} - F1_{in}^{OO}}{F1_{in}^{OO}} + \frac{F1_{ex}^{ij} - F1_{ex}^{OO}}{F1_{ex}^{OO}} \right) \times 100\%, \qquad (1)$$

where $\delta F1^{ij} = F1_{in}^{ij} - F1_{ex}^{ij}$ is the F1 score difference between internal and external validations, $i \in \{O, U, Over, E\}$ is the index for the chosen sampling strategy including the original data without a sampling strategy (O), under-sampling (U), over-sampling (Over) and ensemble method (E), respectively; $j \in \{DR, O\}$ indicates if the dimensional reduction technique is used (DR) or not (O).

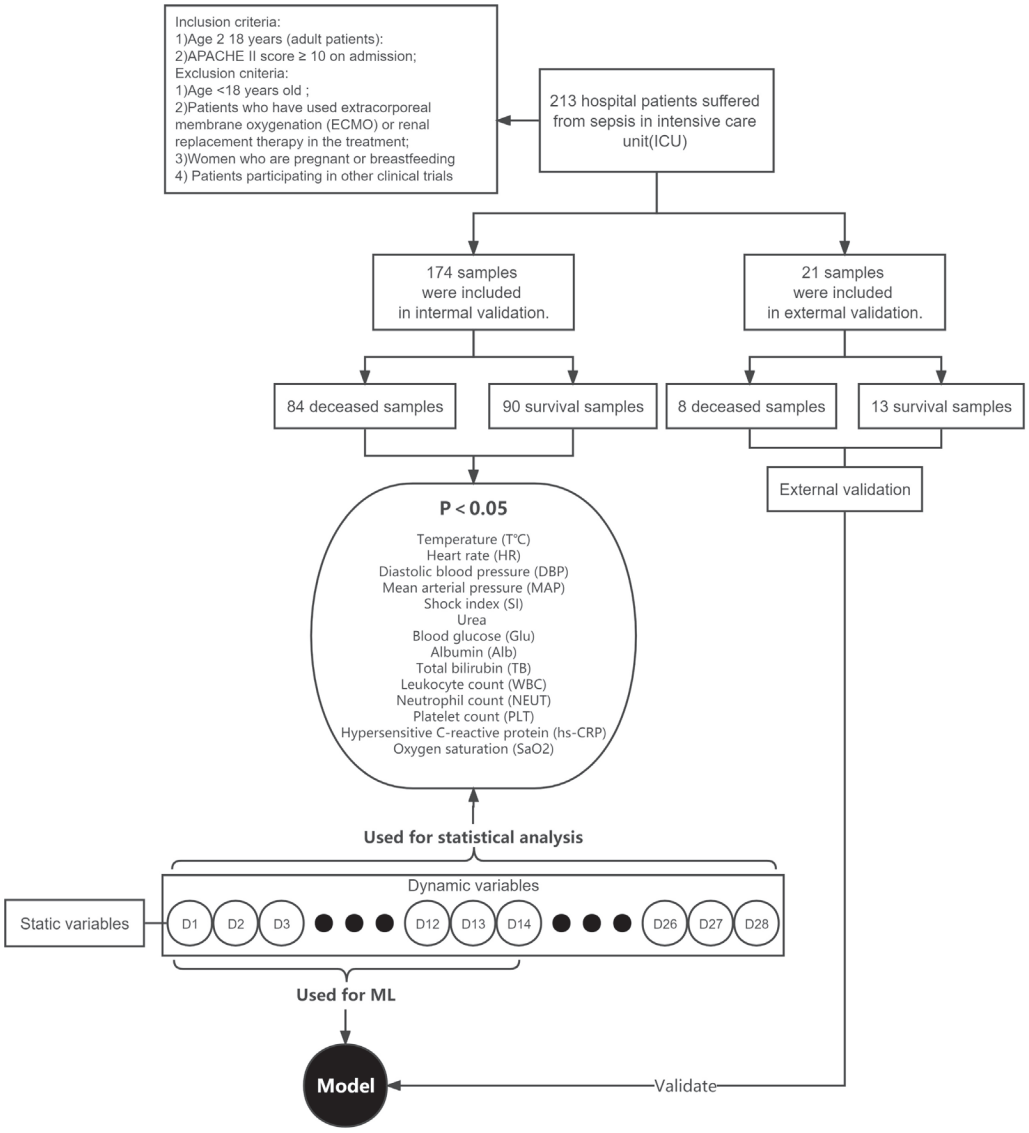

Fig. 1. Comprehensive flowchart illustrating the data acquisition pipeline and subsequent statistical analysis processes.

2.4. ML-input data engineering

2.4.1. Concatenated-triplet data structures for ML-input

To encode the temporal feature of the longitudinal dataset, we use concatenated triplets to create the ML-ready input data for machine learning in our modeling approach. Let $\mathbf{x}_i$ denote the standardized dynamic indicators of a patient at the ith day, and y the corresponding clinical outcome (survival or mortality/deceased) on the 14th day. The ML-ready data vector is engineered by concatenating three components, normalized static indicators $\mathbf{x}_{static}$, consecutive dynamic features of k days $\mathbf{x}_{dynamic} = (\mathbf{x}_i, \mathbf{x}_{i+1}..., \mathbf{x}_{i+k-1})$, and the label y. We call this a "concatenated triplet" and denote it as $(\mathbf{x}_{static}, \mathbf{x}_{dynamic}, y)$. By concatenating dynamical features in k consecutive days, we effectively encode the temporal feature in the longitudinal data into the ML-ready dataset. From the longitudinal record of one patient in T days, we can generate $T - k + 1$ concatenated triplets by sliding the window with width k over the dynamic features of T-day's data as shown in Fig. 2. This allows us to greatly increase the size of the patients' dataset by at least $T - k + 1$ folds (see A.2).

In our dataset, there are 117 surviving subjects (negative class) and 57 deceased subjects (positive class), as shown in A.2. The ratio of the positive to negative class is approximately 0.5, indicating a class imbalance in the classification problem. The standard classifiers designed for balanced datasets always have a bias on the majority class of imbalanced datasets, which can lead to less reliable results when applied to imbalanced datasets [9,10]. Meanwhile, for a given high-dimensional dataset, the higher the dimension is, the sparser the data points are, which may prevent standard classifiers designed for low-dimensional spaces from correctly classifying the points in high dimensional spaces [14]. It is therefore crucial to remove the effect of class imbalance and to mitigate the issue of high dimensionality. Sampling strategies and dimension-reduction techniques are powerful tools to tackle these problems.

In this paper, we propose a framework to deal with the challenge by organically combining sampling strategies and dimension-reduction techniques with standard classifiers. In the framework shown in Fig. 3, we first select the important dynamic features by dimension reduction method and then generate the concatenated triplets. Afterward, ML-ready dataset containing all concatenated triplets is randomly split into the training set and test set for machine learning. An additional dataset is used for external validation to assess the clinical performance of the classifiers.

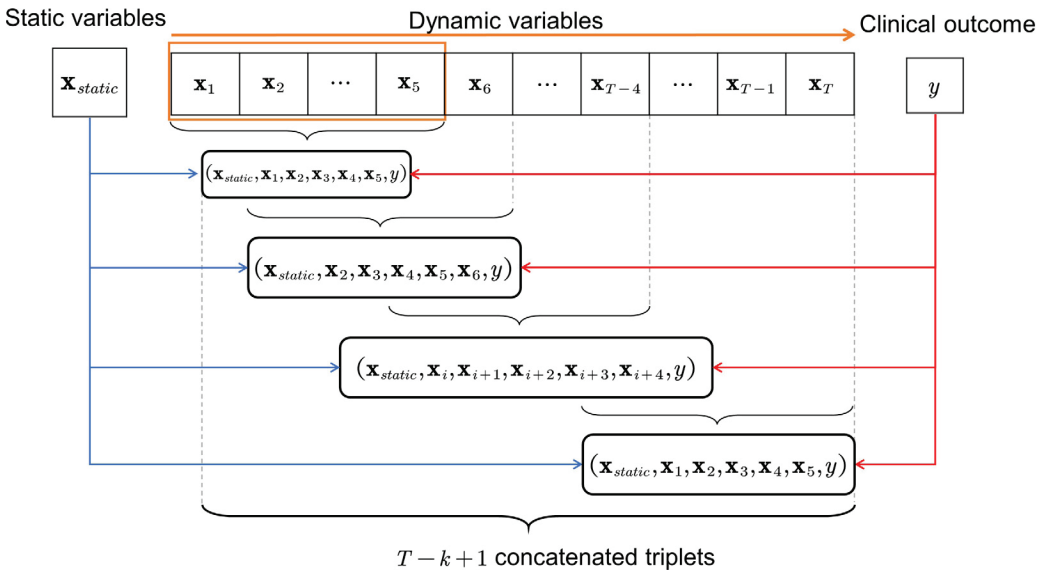

Fig. 2. Schematic representation of the concatenated triplet generation process. This diagram illustrates the method of generating T-k+1 concatenated triplets from a patient's longitudinal record spanning T days. A sliding window of width k=5 is employed to process the temporal data. Each triplet encapsulates a distinct 5-day period, with successive triplets overlapping by 4 days. The label y represents the patient's clinical outcome on the 14th day post-admission, serving as the target variable for predictive modeling. Key elements: (a) T: Total number of days in the patient's record; (b) k: Width of the sliding window (set to 5 days); (c) T-k+1: Total number of generated triplets; (d) y: Binary outcome label (e.g., survival status) at day 14. This data structuring approach enables the capture of temporal patterns while significantly augmenting the dataset size, facilitating more robust machine learning model training.

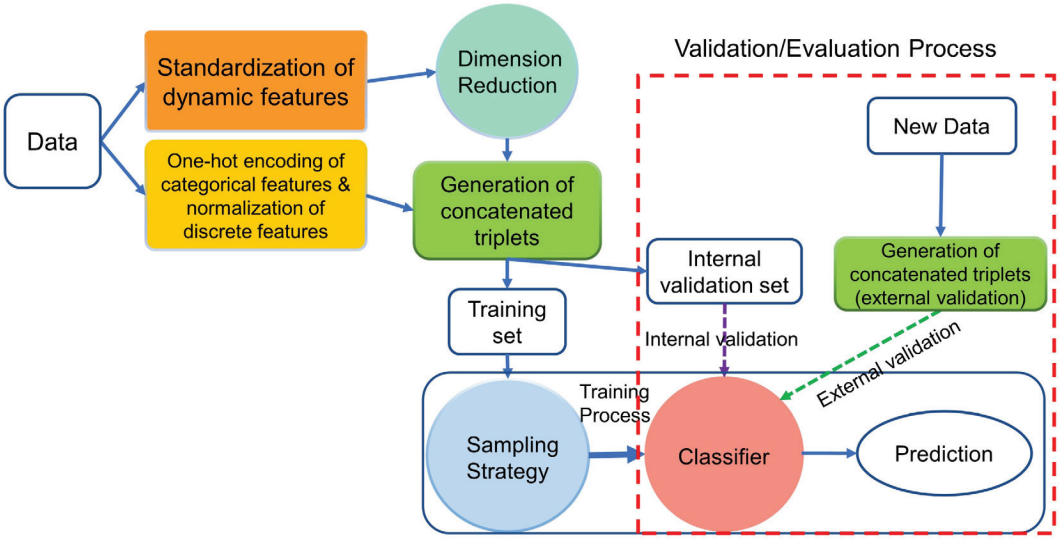

Fig. 3. Comprehensive diagram of the proposed machine learning framework for small dataset challenges in clinical prognosis. This schematic illustrates our novel approach to handling small, high-dimensional clinical datasets: (1) Dimension Reduction: Standardized high-dimensional dynamic features undergo dimension reduction to mitigate the curse of dimensionality. (2) Concatenated Triplet Generation: The reduced-dimension data is used to create concatenated triplets, encoding temporal patterns and augmenting the dataset. (3) ML-Ready Dataset: All concatenated triplets from patient records are compiled into an ML-ready dataset. (4) Data Splitting: The dataset is randomly partitioned into training and test sets. (5) Model Training: Classifiers are trained on the training set using appropriate sampling strategies to address class imbalance. (6) Validation: Well-trained classifiers undergo validation on both the held-out test set and new, unseen data to assess generalizability. This framework addresses key challenges in clinical machine learning, including high dimensionality, temporal dependencies, and limited sample sizes, while ensuring robust model performance and generalizability.

2.4.2. *Sampling strategies*

To address the class imbalance issue, one commonly adopts two approaches: one at the data level while the other at the algorithmic level, leading to specialized classifiers [10]. We adopt the former in this study encompassing the four strategies listed below.

1. Original case (O): No sampling strategy is applied which is treated as the baseline.
2. Random under-sampling strategy (U): randomly under-sample strategy is applied to the majority class to match the size of the minority class.
3. Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC): the minority class is over-sampled to match the majority class using k nearest neighbors A.1 [15].
4. Ensemble under-sampling strategy (E): One splits the majority class into smaller groups and trains multiple classifiers for each

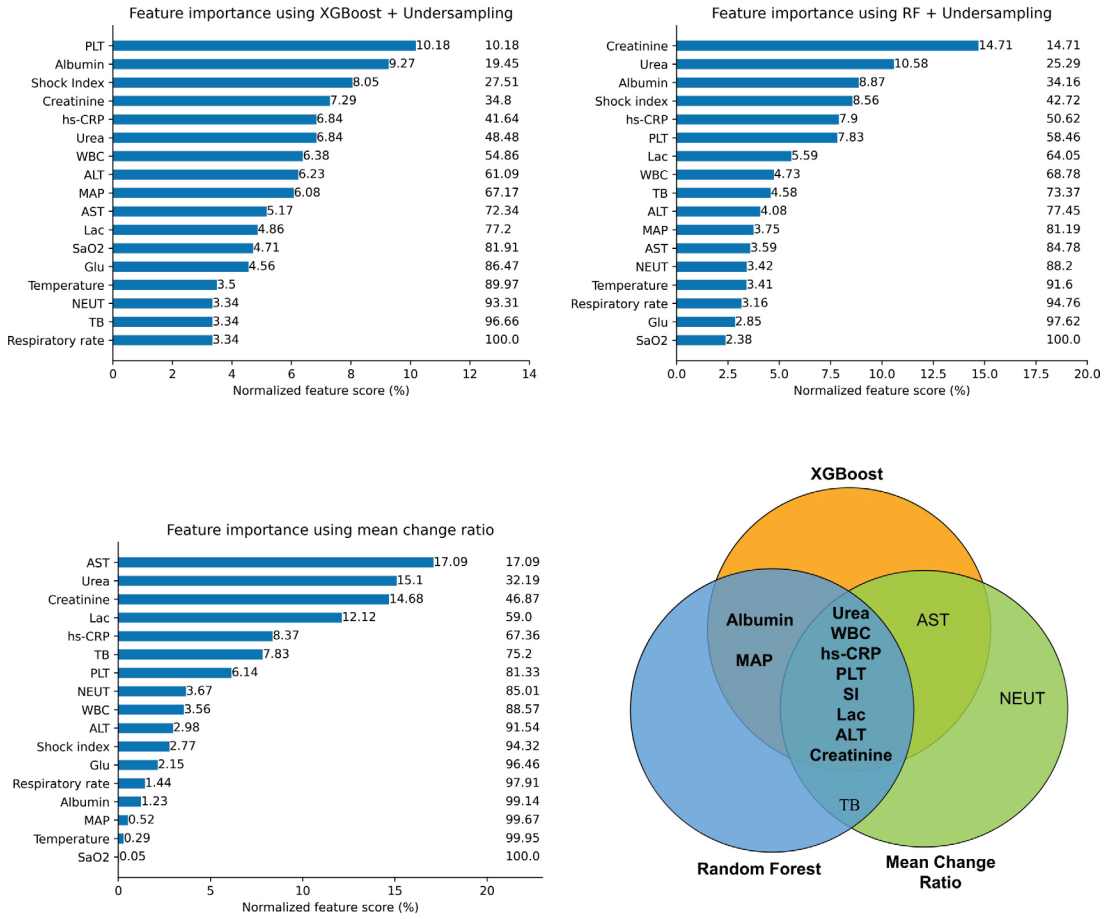

Fig. 4. Multi-panel analysis of feature importance and selection across different methods. This figure presents a comprehensive analysis of feature importance and selection using multiple approaches: (a) Top left: Feature importance ranking derived from XGBoost. (b) Top right: Feature importance ranking derived from Random Forest (RF). (c) Bottom left: Mean change ratio of features. (d) Bottom right: Venn diagram illustrating the relationships among top-ranked features. The key elements include: (1) Cumulative scores: The right-side numbers in panels (a), (b), and (c) represent the cumulative importance scores for features, listed from top to bottom. (2) Feature selection criteria: The Venn diagram displays the overlap among feature sets with cumulative scores exceeding 75% from XGBoost, RF, and mean change ratio analyses. (3) Final feature set: The top 10 features common to both XGBoost and RF results are selected as the final feature set. (4) Abbreviations: A comprehensive table of indicator abbreviations is provided in Appendix A.1. This multi-method approach ensures robust feature selection by leveraging the strengths of different machine learning algorithms and statistical measures.

set consisting of a group of majority class and the whole minority class to get a final ensemble model using the majority voting A.2 [16].

2.4.3. Dimension reduction techniques

Dimension reduction can be accomplished by either the feature selection or feature extraction method [17]. To keep the model interpretable, we use the feature selection method. Before applying the method, the under-sampling strategy is applied to 17 dynamic features on the 14th day to eliminate the effect of class imbalance and avoid introducing artificial noises. To ensure the quality and robustness of the selected features, two methods, RF and XGBoost, for selecting features based on different criteria A.5, are used to identify important features. The common features of two sets of top features obtained separately from RF and XGBoost with cumulative contributions of $\geq$ 75% are selected as the final features to represent the longitudinal data. To further check the quality of the selected features, we define a metric called the mean change ratio, based on the severity of the indicator changes as follows

$$\text{mean change ratio} = \frac{|\mu_2 - \mu_1|}{\mu_2} + \frac{|\mu_2 - \mu_1|}{\mu_1}, \tag{2}$$

where μ_1, μ_2 are the means of the features of dead and alive samples, respectively. Intuitively, the larger the mean change ratio is, the more important the feature should be.

Ranking the feature importance as shown in Fig. 4 by three methods: RF, XGBoost, and mean change ratio, in which 11 top features ranked by each of the three methods are displayed. There are 8 common features among the feature selection results based on RF, XGBoost, and the mean change ratio as the Venn diagram shows in Fig. 4. As Albumin and MAP have relatively high contributions in XGBost and RF, we choose the 8 common features augmented by Albumin and MAP to arrive at the 10 final features with a cumulative contribution exceeding 75%. They are Urea, WBC, hs-CRP, PLT, SI, Creatinine, Lac, ALT, Albumin, and MAP, respectively.

3. Results and discussion

3.1. Results

The ML-ready dataset is randomly split into a training set and a test set with a ratio of 8:2 for training and testing (or internal validation) in machine learning, respectively. We implement five classifiers: K Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). XGBoost is implemented using the XGBoost package (version 1.6.) [18]. KNN, LR, RF, and SVM are implemented using scikit-learn (version 1.1.1). Three different sampling strategies are applied to the training set using the imbalanced-learn package (version 0.7.0) [19]. The hyperparameters of the classifiers are selected using a grid search method with 5-fold cross-validation on the training set, given in Appendix A.

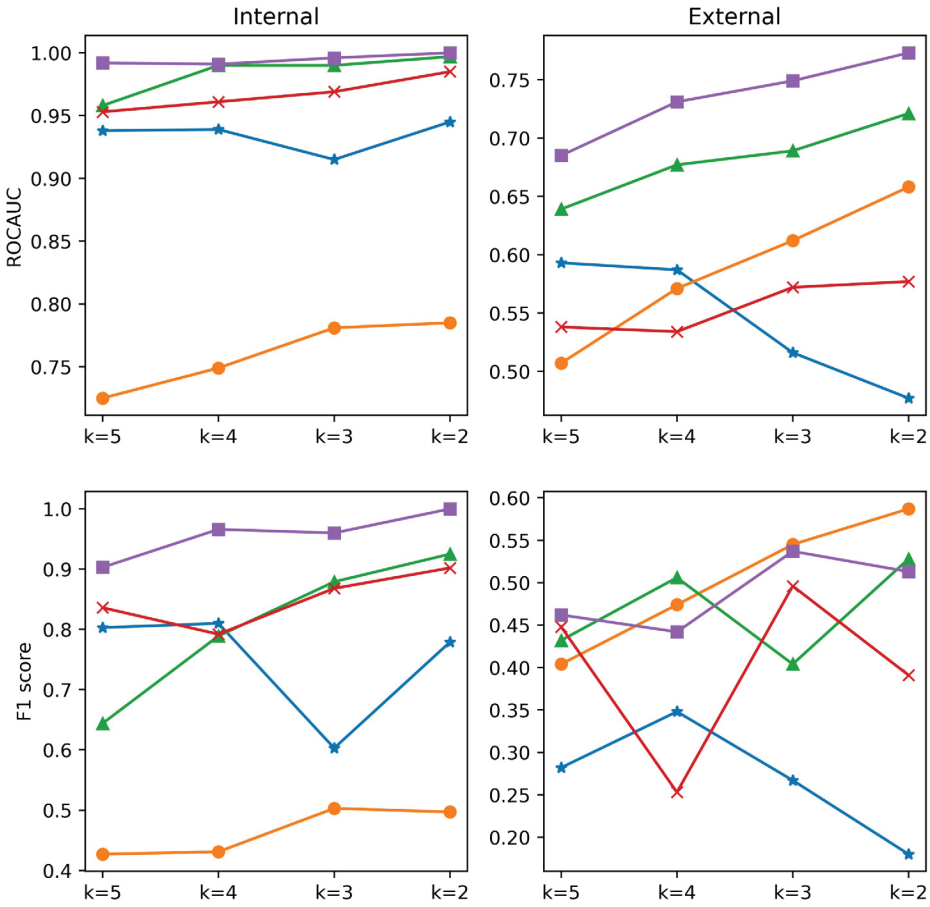

Fig. 5. Baseline performance of five classifiers, without feature engineering, is presented in terms of AUROC (top row) and F1 score (bottom row). The term "baseline" refers to the use of the original dataset without applying any sampling strategies or dimensionality reduction techniques, across various k values, for training the classifiers. This comparison underscores the critical role of feature engineering in mitigating the risk of overfitting and enhancing model accuracy. The first row displays AUROC values, while the second row illustrates F1 scores for both internal validation (left) and external validation (right). The classifiers are represented as follows: KNN with stars (blue), LR with circles (orange), RF with triangles (green), SVM with crosses (red), and XGBoost with squares (purple). Notably, LR demonstrates the most consistent F1 scores in both internal and external validations, indicating a low risk of overfitting. In contrast, the inconsistency in F1 scores between internal and external validations for the remaining four classifiers suggests a higher risk of overfitting. Specifically, the larger the difference in F1 scores between internal and external validations, the higher the likelihood that the model is overfitting to the training data and failing to generalize to new, unseen data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We first present the baseline results of the models without feature engineering in terms of the AUROC and F1 score in Fig. 5. We observe that (a) as validated both internally and externally, the smaller the k value is, the higher the value of AUROC for all classifiers except KNN. It indicates that all classifiers except KNN perform better for small k; (b) based on AUROC, XGBoost and RF outperform the others in both the internal and external validation; (c) XGBoost (AUROC = 0.773, F1 score = 0.513 at k = 2, AUROC = 0.685, F1 score = 0.462 at k = 5) outperforms RF (AUROC = 0.721, F1 score = 0.528 at k = 2, AUROC = 0.639, F1 score = 0.432 at k = 5) in the external validation; (d) all classifiers except for LR are at a high risk of overfitting after comparing the internal with the external validation results in both the AUROC and F1 score.

We then examine the efficacy of the sampling strategies and plot the AUROC for the external validation with 3 different sampling strategies and k values in Fig. 6. Comparing the results of the three sampling strategies with the baseline cases, we find that AUROC values are not altered much (a decrease at most 0.126 and an increase at most 0.058 with under-sampling, a decrease at most 0.068 and an increase at most 0.119 with over-sampling, a decrease at most 0.041 and an increase at most 0.033 with the ensemble method). This indicates that the AUROC is not sensitive to various sampling strategies at all.

We then examine the degree of improvement in overfitting γ in Fig. 7. Among the sampling strategies, the ensemble method works the best for the classifiers in most cases except for KNN at k = 5, SVM is the most sensitive to various sampling strategies while RF is the least sensitive. For KNN, the over-sampling strategy always works for all k values. This might be because the over-sampling algorithm SMOTE-NC uses a randomly selected k nearest neighbor of an original sample to generate a new sample. For RF, the ensemble under-sampling works for all k values. For SVM, over-sampling never works while, for XGBoost, ensemble under-sampling always works the best for most k values.

Analysis of Fig. 7 reveals several key observations:

(1) At k=5, there are 6 bars with negative values and 5 with positive values.
(2) As k decreases to 2, the number of bars with negative values progressively diminishes.
(3) Specifically, k=3 and k=2 yield one and two negative values, respectively.
(4) Across all k values and classifiers, at least one bar consistently displays a positive value.

These observations suggest that the degree of overfitting for all classifiers can be mitigated at any k value through the application of appropriate sampling strategies. Consequently, we conclude that judiciously chosen sampling strategies, when combined with standard classifiers, can enhance the models' generalizability without significantly compromising accuracy (as measured by AUROC).

Dimension reduction techniques combined with various sampling strategies are implemented for k = 2 since most models perform better

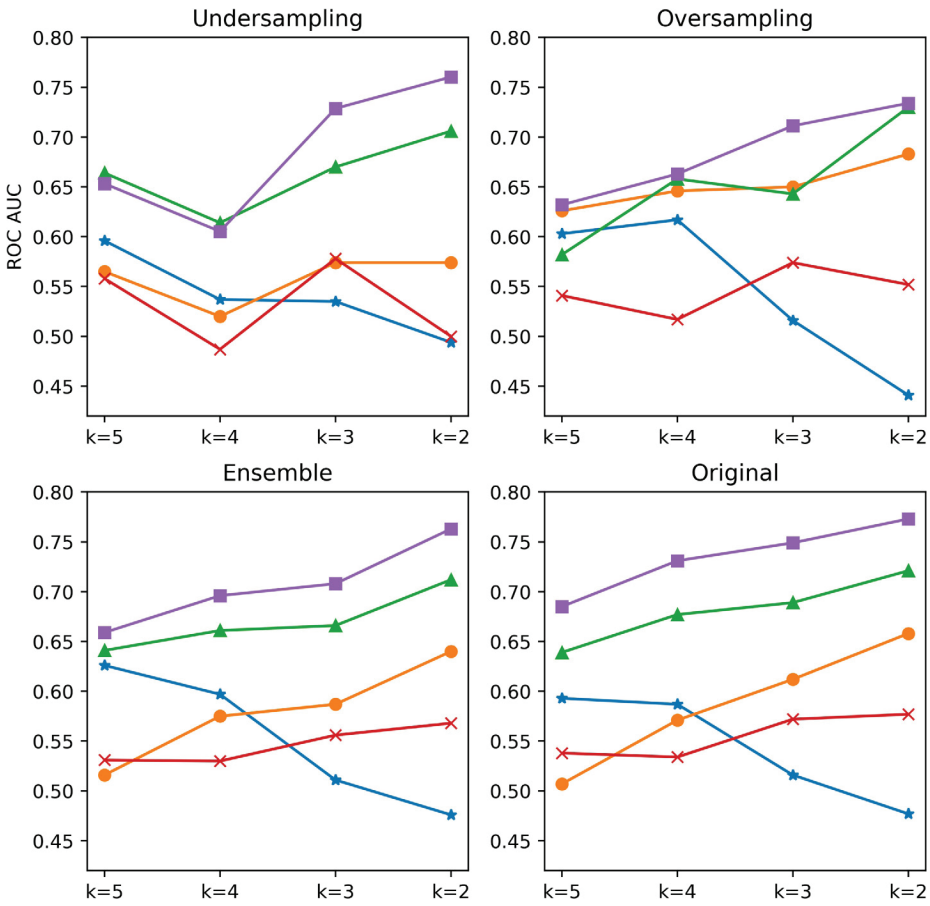

Fig. 6. AUROC values from the external validation results for three different sampling strategies, compared with the baseline (original data) across various k values. The plots are organized as follows: top left for under-sampling, top right for oversampling, bottom left for ensemble under-sampling, and bottom right for the baseline. The classifiers are represented as follows: KNN with stars (blue), LR with circles (orange), RF with triangles (green), SVM with crosses (red), and XGBoost with squares (purple). By comparing these results with the baseline, it becomes apparent that the AUROC metric exhibits minimal sensitivity to the different sampling strategies, as the AUROC values remain largely consistent with the baseline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on the short-term stacked concatenated triplets. The results in the external validation are shown in Fig. 8 and Appendix A. Comparing the results among the classifiers built on selected features (O+DR) and the original features (O+O) as shown in Fig. 8, we find that the AUROC values are improved for all classifiers, while the F1 scores are reduced in most cases. More precisely, the improvement in the AUROC value is 0.226 (KNN), 0.075 (LR), 0.081 (RF), 0.086 (SVM), and 0.030 (XGBoost), respectively. Similarly, the improvement in the F1 score is 0.286 (KNN), 0.057 (LR), −0.003 (RF), −0.391 (SVM), −0.056 (XGBoost), respectively. So, dimensionality reduction techniques can improve model accuracy, but the resulting models may be more prone to overfitting, suggesting that sampling strategies need to be used.

Considering different classifiers, we find that XGBoost and RF outperform their peers in all cases and XGBoost works better than RF except for the under-sampling case (U+DR) in the external validation. Comparing 3 sampling strategies, XGBoost (AUROC = 0.777, F1 score = 0.694) and RF (AUROC = 0.769, F1 score = 0.647) achieve the best performance in the E+DR case. However, both XGBoost (AUROC = 0.803, F1 score = 0.457) and RF (AUROC = 0.802, F1 score = 0.525) achieve better AUROC values but lower F1 scores in the baseline case (O+DR), indicating that there is a trade-off between the accuracy and generalization error in the models when the F1 score is used as a metric for the generalization error.

To show the effectiveness of the proposed modeling framework, we compare the improvement in overfitting and performance metrics in the three cases with the baseline case as shown in Fig. 9. Based on the F1 score, we find that when only dimension reduction techniques are used for the class imbalanced data, the degree of overfitting is improved only in KNN since all other diamonds lie below the baseline (zero) in Fig. 9(a). However, when one combines dimension reduction techniques with sampling strategies, overfitting in all trained classifiers except for SVM is greatly improved, which demonstrates the importance of proper sampling strategies on the class-imbalanced problem. Moreover, the results show that different sampling strategies have different effects on different classifiers. The ensemble under-sampling strategy (E) is the best strategy for KNN, RF and XGBoost compared to their counterparts with the improvement in overfitting (314.75%, 54.96% and 77.72%, respectively). The least improvement shown in KNN, LR, RF, XGBoost is 176.48% (Over), 102.63% (U), 25.30% (Over) and 17.29% (Over).

An analysis of AUROC shows that the highest and lowest improvement of AUROC are 53.46% (E) and 27.46% (Over) for KNN, 11.55% (O) and 5.78% (U) for LR, 11.23% (O) and 1.25% (Over) for RF, 18.20 (Over) and −2.95% (U) for SVM, and 3.88% (O) and −7.76% (U) for XGBoost, respectively. From Fig. 9(b), we notice that the proposed framework can greatly improve the AUROC value. But the

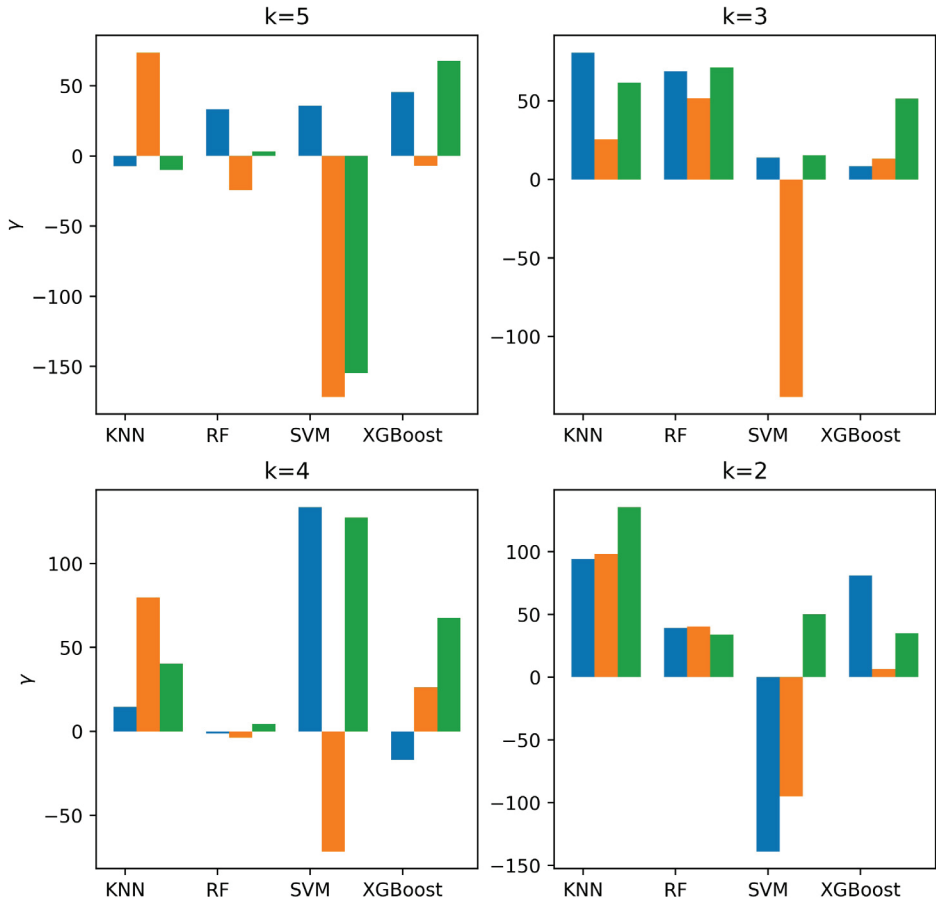

Fig. 7. Improvement in overfitting, measured by γ, for four classifiers across four k values. In each subplot, the bars represent different sampling strategies: the left (blue) bar for under-sampling, the middle (orange) bar for oversampling, and the right (green) bar for ensemble under-sampling. Bars above the zero line indicate an improvement in overfitting compared to the baseline (original data without sampling). The higher the bar above the zero line, the greater the improvement in overfitting. Notably, for each classifier and across all k values, there is always at least one bar above the zero line, suggesting that there is always a suitable sampling strategy to mitigate overfitting for each classifier. Since LR demonstrated the most consistent F1 scores in both internal and external validations (as shown in Fig. 5), indicating a low risk of overfitting, it has been excluded from these plots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

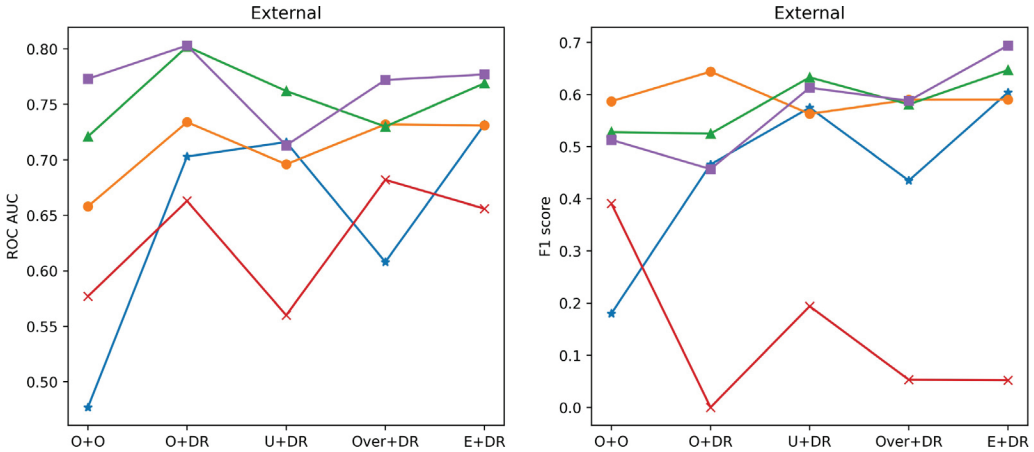

Fig. 8. The AUROC values (left) and F1 scores (right) from the external validation of dimensional reduced models, combined with three different sampling strategies and the baseline, are presented for five classifiers to highlight the effectiveness of feature engineering. Each line, distinguished by color and marker, represents a well-trained classifier: KNN is marked by stars (blue), LR by circles (orange), RF by triangles (green), SVM by crosses (red), and XGBoost by squares (purple). The x-axis labels are as follows: "O" denotes the original data without any sampling or dimensionality reduction, with "O+O" representing the baseline; "DR" stands for Dimensionality Reduction; "U" for random under-sampling; "Over" for oversampling (SMOTE-NC); and "E" for Ensemble under-sampling. The notation 'A+B' is used to indicate the combination of a sampling strategy with dimensionality reduction. For example, "O+DR" represents the original data combined with dimensionality reduction, and "U+DR" denotes under-sampling combined with dimensionality reduction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

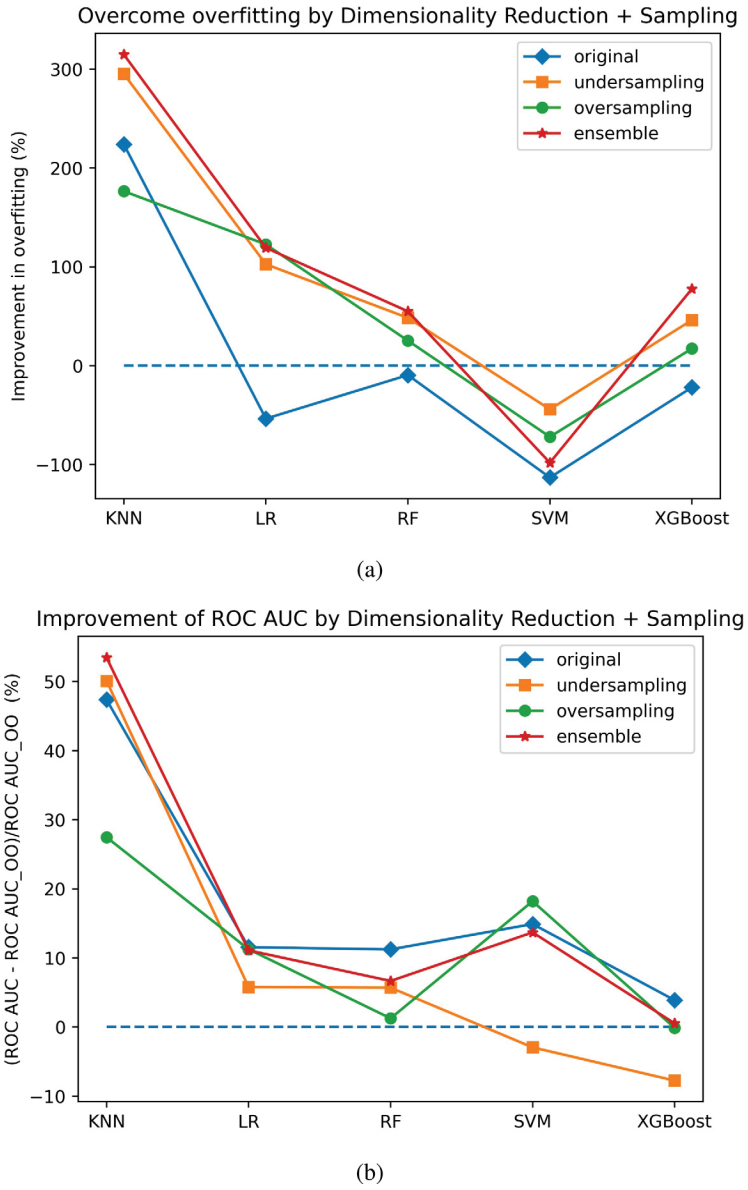

Fig. 9. A comparative analysis of feature engineering, combining dimensionality reduction with sampling strategies, against the baseline. The upper panel illustrates the improvement in mitigating overfitting, quantified by the percentage of improvement. The lower panel demonstrates the enhancement in model performance, measured by the Area Under the Receiver Operating Characteristic (AUROC) curve. This dual representation provides a comprehensive assessment of how dimensionality reduction techniques and sampling strategies impact both the model's generalizability (improvement in overfitting) and its predictive accuracy (improvement in AUROC).

under-sampling strategy slightly reduces the AUROC value to −2.95%, −7.76% for SVM and XGBoost, respectively.

When examining Fig. 9(a) and (b) together, we notice that KNN benefits the most, and XGBoost and RF benefit the least from feature engineering compared to others. This is because KNN is more prone to overfitting while XGBoost and RF have several mechanisms to prevent overfitting. SVM is a special case in the sense that the accuracy is improved but the overfitting is aggravated. This suggests that other techniques at the algorithm level can be beneficial for SVM.

In summary, the proposed modeling framework works well for the small data problem with high dimensional data and imbalanced classes; the sampling strategy can mitigate overfitting while slightly sacrificing the accuracy of some classifiers. Hence, to achieve the accuracy and generalizability of the model, one needs to carefully choose the appropriate classifier and sampling strategy combo and be fully aware of the trade-off between accuracy and generalizability.

3.2. Discussion

The challenge of working with small datasets is a persistent issue in data science, particularly in medical research [7]. While previous studies in this field have primarily focused on small-sized, high-dimensional, but class-balanced datasets [8], our study breaks new ground. To the best of our knowledge, this is the first successful attempt to construct prognostic models based entirely on small-sized longitudinal data that simultaneously exhibits class imbalance and high dimensionality in clinical medicine research.

Our proposed modeling framework demonstrates its effectiveness by developing prognostic models for sepsis that combine advanced classifiers with sophisticated feature engineering techniques. The results are promising: XGBoost, when coupled with an ensemble under-sampling strategy and carefully selected important features, achieves the best performance with an AUROC of 0.777, an F1 score of 0.694, and a low generalization error.

A key innovation in this study is the introduction of a novel input data structure, which we term the "concatenated triplet." This structure serves a dual purpose:

- It effectively encodes the temporal characteristics of the longitudinal data.
- It augments the size of the input dataset, addressing the small data challenge.

The efficacy of this ML-ready dataset structure is corroborated by our results, underscoring its potential for similar applications in medical data analysis.

Class imbalance in datasets significantly impedes the performance of standard learning algorithms, often leading to their failure in generalizing inductive rules across the sample space and resulting in overfitting [10]. Overfitting has been a source of serious errors in model predictions and remains one of the principal challenges in machine learning [20–22]. However, this issue is not always apparent from commonly used performance metrics, and the bias towards the majority class is frequently overlooked.

To address this, we introduce a new metric, γ, derived from the F1 score, to quantify improvements in mitigating overfitting. While previous studies in 2019 and 2021 explored class imbalance in sepsis data [23,24], they primarily focused on assigning weights to important features—an approach that has shown limited potential for substantial performance improvements.

Our study takes a different approach, employing three distinct sampling strategies to tackle the class imbalance issue:

- Random under-sampling
- SMOTE-NC over-sampling
- Ensemble under-sampling

Our findings demonstrate that pairing an appropriate classifier with a suitable sampling strategy can significantly mitigate overfitting. Specifically, we observed improvements of 54.96% with XGBoost and 77.72% with Random Forest (RF).

Several studies have explored sampling strategies to mitigate the effects of imbalanced datasets in sepsis research. Adam Karlsson et al. (2021) employed an under-sampling strategy [25], achieving AUROC values of 0.83 and 0.80 for 7-day and 30-day predictions, respectively, in internal validation. However, their study lacked external validation and F1 score reporting, limiting comprehensive performance assessment.

Recent literature suggests under-sampling as a superior strategy [26,27], as demonstrated in a sepsis prognostic study [28]. Our research advances this understanding by showing that ensemble under-sampling is most effective. This approach inherits under-sampling's advantages while preserving all original dataset information, making it particularly adept at addressing class imbalance. We found that different sampling strategies impact classifiers variably, potentially explaining why a 2021 study [24] reached no firm conclusions. Our results demonstrate that appropriate classifier-sampling strategy pairing significantly reduces overfitting caused by class imbalance. Thus, our proposed modeling approach (combining standard classifiers with proper sampling strategies) effectively addresses imbalanced data issues.

Septic patient data in high-dimensional space is typically sparse and edge-distributed [14], potentially hindering standard classifiers. Selected low-dimensional data representations can effectively denoise while retaining crucial features, enabling simpler, more comprehensible models [17,29]. Dimension reduction thus serves as an effective means of denoising, feature selection, and simplification, particularly useful for biological datasets [30].

In our study, XGBoost and Random Forest, combined with under-sampling, address issues caused by high-dimensional, imbalanced data. We enhance feature selection credibility using mean change ratios of dynamic features. Our results show that feature selection effectively improves clinical prognosis accuracy (AUROC) for septic patients [17,29,31]. Moreover, properly combining sampling strategies with dimension reduction techniques significantly enhances both accuracy and generalizability of standard classifiers.

In summary, our proposed modeling framework performs well on small datasets, such as sepsis prognosis. Optimal performance and generalizability require careful selection of classifier-sampling strategy pairs and balancing accuracy against generalizability (overfitting). However, limitations exist: this framework may not suit all standard classifiers (e.g., SVM), and other algorithmic-level techniques could further improve the model. The approach's effectiveness requires verification on additional clinical medicine problems.

4. Conclusion

In this study, we have successfully developed an innovative machine learning framework for creating prognostic tools for septic patient's outcomes in a given temporal window, effectively addressing the challenges posed by clinically available datasets that are small, high-dimensional, and imbalanced. Our approach introduces several key advancements:

- Data Augmentation: We engineered a novel data structure, the "concatenated triplet", which significantly expands the effective size of the dataset while encoding crucial temporal features of the longitudinal data.
- Superior Performance: Our best models demonstrate performance that surpasses externally validated machine learning models developed on substantially larger datasets, underscoring the efficacy of our approach.
- Feature Engineering: By combining standard classifiers with sophisticated feature engineering techniques, our modeling framework showcases its power in developing robust prognostic tools for clinical applications.
- Generalizability: The success of this framework in handling small, complex datasets opens new avenues for machine learning applications in the medical field, where large, well-balanced datasets are often unavailable.
- Clinical Relevance: Our approach bridges the gap between advanced machine learning techniques and practical clinical needs, offering a viable solution for developing accurate prognostic tools with limited data resources.

This framework not only addresses the immediate challenge of sepsis prognosis but also provides a blueprint for tackling similar issues across various medical domains. By demonstrating that sophisticated machine learning approaches can yield high-performance models even with limited data, we pave the way for broader adoption of AI-driven decision support tools in clinical settings where data scarcity has traditionally been a barrier.

Future research directions may include: (1). extending this framework to other medical conditions with similar data challenges, (2). exploring the integration of this approach with existing clinical decision support systems, (3). investigating the potential for real-time application of these models in clinical settings. Our study represents a significant step forward in applying machine learning to complex medical problems, offering a promising pathway for developing accurate, clinically relevant prognostic tools even in data-constrained environments.

CRediT authorship contribution statement

Chunyan Li: Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Lu Wang:** Conceptualization, Data curation, Formal analysis, Investigation, Validation, Writing – original draft. **Kexun Li:** Conceptualization. **Hongfei Deng:** Data curation. **Yu Wang:** Data curation. **Li Chang:** Data curation. **Ping Zhou:**

Data curation. **Jun Zeng:** Project administration. **Mingwei Sun:** Project administration. **Hua Jiang:** Conceptualization, Data curation, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. **Qi Wang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Statements of ethical approval

This retrospective study was approved by the Ethics Committee of Sichuan Provincial People's Hospital (No. 30 in 2021) and has been registered in the China Clinical Trial Registration Center (Clinical Registration No.: ChiCTR2100050799). The requirement for written informed consent was waived due to the retrospective nature of the study.

Declaration of competing interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The research of Chunyan Li and Qi Wang is partially supported by NSF award DMS-1954532, OIA-2242812, and an SC EPSCOR GEAR award. Hua Jiang's research is supported by funding from the Sichuan Science and Technology Program, Grant /Award Number: 2021YFS0378.

Appendix A

A.1. List of dynamic indicators and abbreviations

In Table 1, we list the 17 dynamic indicators and their abbreviations, among which the top 10 (starred ones) are eventually selected as important features to be used in the classification tasks. In Table 2, we list all the static indicators used in the study.

A.2. Augmentation of the dataset by concatenated triplets

We have 174 patients' records in 14 days in total. The number of deceased patients each day within 2 weeks is listed in Table 3. Since we have smaller deceased samples, so we call deceased samples positive samples and label them by 1. Then, we have 117 negative samples and 57 positive samples. The numbers of structured concatenated triples with respect to k=2,3,4,5 are shown in Table 4.

A.3. SMOTE-NC

The pseudocode of synthetic minority over-sampling technique which is designed for continuous features is summarized in Algorithm A.1. SMOTE-NC slightly changes the way a new sample is generated by performing something specifically for the categorical features. In fact, the categories of a newly generated sample are decided by picking the most frequent category of the nearest neighbors present during the generation.

Algorithm A.1 (*Algorithm of Synthetic Minority over-sampling Technique*)**.**

Input: positive samples P, negative samples N, $|P| < |N|$, number of nearest neighbours k. Output: Synthetic Minority class samples.

1: ***for*** $i \leftarrow 1, ..., |P|$ ***do***
2: *Compute k nearest neighbours for x_i.*
3: ***end for***
4: *Randomly choose k nearest neighbours of x_i, call it x_{zi}.*
5: *Compute the new synthetic sample $x_{new} = x_i + \lambda(x_i - x_{zi})$ with λ being a random number in $[0, 1]$.*

Table 1
List of total 17 dynamic indicators and abbreviations and the top 10 chosen ones (*).

Indicators	Abbreviations
Urea (mmol/L)	Urea*
White blood cell count ($\times 10^9$/L)	WBC*
High-sensitivity C-reactive protein (mg/L)	hs-CRP*
Platelet Count ($\times 10^9$/L)	PLT*
Shock Index (bmp/mmHg)	SI*
Creatinine (μmol/L)	Creatinine*
Lactic acid (mmol/L)	Lac*
Alanine aminotransferase (U/L)	ALT*
Albumin (g/L)	Albumin*
Mean arterial pressure (mmHg)	MAP*
Total bilirubin (μmol/L)	TB
Temperature (C)	Temperture
Respiratory rate (bmp)	Respiratory rate
Blood glucose (mmol/L)	Glu
Aspartate aminotransferase (U/L)	AST
Neutrophil count ($\times 10^9$/L)	NEUT
Oxygen saturation (%)	SaO_2

Table 2
The list of 9 static indicators.

Indicators	Feature Type
Age	Categorical variable
Gender	Categorical variable
Diagnosis	Categorical variable
Mental status at admission	Categorical variable
Admitted for sepsis or not	Categorical variable
PN+EN or not	Categorical variable
APACAE II Score	Non-categorical discrete variable
SOFA Score	Non-categorical discrete variable
BMI	Non-categorical discrete variable

Table 3
Number of deceased patients in each day.

Day	5	6	7	8	9	10	11	12	13	14	Total
# of deceased patients	1	1	5	5	6	6	3	13	4	13	57

Table 4
Number of ML-ready data with the structured concatenated triplets for various k values generated from 174 patients consisting of 117 survival patients and 57 deceased patients at the end of the 14th day. The last column is the ratio of total number of ML-ready data to total number of patients.

	Survival (negative samples)	Deceased (positive samples)	Total samples	Ratio
$k = 5$	1170	395	1565	8.99
$k = 4$	1287	452	1739	9.99
$k = 3$	1404	509	1913	10.99
$k = 2$	1521	566	2087	11.99

A.4. Ensemble under-sampling technique

The pseudocode of the ensemble under-sampling technique can be summarized in Algorithm A.2. One can split the negative samples into subsets such that $|N_i| \sim |P|$ (almost the same size) or strictly $|N_1| < |P|$ so that the negative class will be the minority class and might impact the model performance. It is hard to say which one works better, depending on the applications. In our case, after several test experiments, we choose $T = 2$ such that $|N_i| \sim |P|$.

Algorithm A.2 (*Algorithm of Ensemble under-sampling Technique*)**.** *Input: All positive samples P, all negative samples N, $|P| < |N|$, number of subsets of negative samples T. Output: prediction result*

1: *Randomly split N into T subsets $\{N_1, N_2, ..., N_T\}$.*
2: ***for*** $i \leftarrow 1, ..., T$ ***do***
3: *Train classifier X using the combined dataset $\{N_i, P\}$ to obtain prediction result R_i.*

*4: **end for***

5: Integrate T results $\{R_1, R_2..., R_T\}$ by majority voting to obtain the final result.

A.5. Computation of feature scores

RF and XGBoost are two feature selection methods that rank the features using two different metrics.

To compute the importance score of each feature in RF, we assume there are totally n_m samples at node m, y is the label of the data, and $I(y = k)$ is the number of data belonging to class k in node or leaf m. Then,

$$p_{mk} = \frac{I(y = k)}{n_m} \tag{3}$$

is the approximation of the probability of class $k \in 0, 1$ observations in node m. Then, the Gini impurity of node m is defined by

$$H_m = \sum_k p_{mk}(1 - p_{mk}). \tag{4}$$

The reduction of impurity of a node m is given by

$$impurity_m = \frac{N_t}{N}\left(H_m - \frac{N_{tR}}{N_t}H_R - \frac{N_{tL}}{N_t}H_L\right), \tag{5}$$

where N is the total number of samples, N_t is the number of samples at the current node m, N_{tL} is the number of samples in the left child, and N_{tR} is the number of samples in the right child. H_m is the impurity of the current node, H_R is the impurity of the right child, H_L is the impurity of the left child. Therefore, the score of a feature f in RF is defined as follows:

$$score_f = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{m\in M_f^{(t)}} impurity_m^{(t)}}{\sum_f \sum_{m\in M_f^{(t)}} impurity_m^{(t)}}\right), \tag{6}$$

where T is the total number of trees in RF, $M_f^{(t)}$ is the collection of nodes that split on feature f in the tth tree.

The importance score of a feature f in XGBoost is defined by

$$score_f = \sum_{t=1}^{T} n_f^{(t)}, \tag{7}$$

where T is the total number of trees in XGBoost, $n_f^{(t)}$ is the number of times each feature f is used to split the data in the tth tree.

To visualize and analyze the feature importance obtained from these two methods in the same scale, we normalize the feature score as follows:

$$\text{normalized score}_f = \frac{score_f}{\sum_f score_f}. \tag{8}$$

References

[1] Rudd Kristina E, Johnson Sarah Charlotte, Agesa Kareha M, Shackelford Katya Anne, Tsoi Derrick, Kievlan Daniel Rhodes, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. Lancet 2020;395(10219):200–11.

[2] Singer Mervyn, Deutschman Clifford S, Seymour Christopher Warren, Shankar-Hari Manu, Annane Djillali, Bauer Michael, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama 2016;315(8):801–10.

[3] 陶亮. "基于泛布尔代数的医疗诊断推理机的原理研究与现实," 博士论文. 武汉理工大学, 2005.

[4] Skyttberg Niclas, Chen Rong, Blomqvist Hans, Koch Sabine. Exploring vital sign data quality in electronic health records with focus on emergency care warning scores. Appl Clin Inf 2017;8(03):880–92.

[5] Forsström JJ, Irjala K, Selén Gustaf, Nyström Mats, Eiuund P. Using data preprocessing and single layer perceptron to analyze laboratory data. Scand J Clin Lab Invest 1995;55(sup222):75–81.

[6] Kokol Peter, Kokol Marko, Zagoranski Sašo. Machine learning on small size samples: A synthetic knowledge synthesis. Sci Prog 2022;105(1):00368504211029777.

[7] Vabalas Andrius, Gowen Emma, Poliakoff Ellen, Casson Alexander J. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14(11):e0224365.

[8] Spiga Ottavia, Cicaloni Vittoria, Fiorini Cosimo, Trezza Alfonso, Visibelli Anna, Millucci Lia, et al. Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. Orphanet J Rare Dis 2020;15(1):1–10.

[9] Prati Ronaldo C, Batista Gustavo EAPA, Monard Maria Carolina. Data mining with imbalanced class distributions: concepts and methods. In: IICAI. 2009, p. 359–76.

[10] He Haibo, Garcia Edwardo A. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21(9):1263–84.

[11] Yang Qiang, Wu Xindong. 10 challenging problems in data mining research. Int J Inf Technol Decis Mak 2006;5(04):597–604.

[12] Rastgoo Mojdeh, Lemaitre Guillaume, Massich Joan, Morel Olivier, Marzani Franck, Garcia Rafael, et al. Tackling the problem of data imbalancing for melanoma classification. In: Bioimaging. 2016.

[13] Van Rijsbergen C. Information retrieval. Dept. Computer Science, Univ. of Glasgow; 1979.

[14] Hastie Trevor, Tibshirani Robert, Friedman Jerome H, Friedman Jerome H. The elements of statistical learning: data mining, inference, and prediction, vol. 2, Springer; 2009.

[15] Chawla Nitesh V, Bowyer Kevin W, Hall Lawrence O, Kegelmeyer W Philip. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[16] Liu Xu-Ying, Wu Jianxin, Zhou Zhi-Hua. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B 2008;39(2):539–50.

[17] Li Jundong, Cheng Kewei, Wang Suhang, Morstatter Fred, Trevino Robert P, Tang Jiliang, et al. Feature selection: A data perspective. ACM Comput Surv 2017;50(6):1–45.

[18] Chen Tianqi, Guestrin Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, p. 785–94.

[19] Lemaître Guillaume, Nogueira Fernando, Aridas Christos K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017;18(1):559–63.

[20] Simon Richard, Radmacher Michael D, Dobbin Kevin, McShane Lisa M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95(1):14–8.

[21] Domingos Pedro. A few useful things to know about machine learning. Commun ACM 2012;55(10):78–87.

[22] Chicco Davide. Ten quick tips for machine learning in computational biology. BioData Min 2017;10(1):1–17.

[23] Reyna Matthew A, Josef Chris, Seyedi Salman, Jeter Russell, Shashikumar Supreeth P, Westover M Brandon, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. In: 2019 computing in cardiology. IEEE; 2019, p. 1.

[24] Misra Debdipto, Avula Venkatesh, Wolk Donna M, Farag Hosam A, Li Jiang, Mehta Yatin B, et al. Early detection of septic shock onset using interpretable machine learners. Jf Clin Med 2021;10(2):301.

[25] Karlsson Adam, Stassen Willem, Loutfi Amy, Wallgren Ulrika, Larsson Eric, Kurland Lisa. Predicting mortality among septic patients presenting to the emergency department–a cross sectional analysis using machine learning. BMC Emerg Med 2021;21(1):1–8.

[26] Mountassir Asmaa, Benbrahim Houda, Berrada Ilham. An empirical study to address the problem of unbalanced data sets in sentiment classification. In: 2012 IEEE international conference on systems, man, and cybernetics. IEEE; 2012, p. 3298–303.

[27] Nieto-del Amor Félix, Prats-Boluda Gema, Garcia-Casado Javier, Diaz-Martinez Alba, Diago-Almela Vicente Jose, Monfort-Ortiz Rogelio, et al. Combination of feature selection and resampling methods to predict preterm birth based on electrohysterographic signals from imbalance data. Sensors 2022;22(14):5098.

[28] Su Longxiang, Xu Zheng, Chang Fengxiang, Ma Yingying, Liu Shengjun, Jiang Huizhen, et al. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. Front Med 2021;8:883.

[29] Nguyen Lan Huong, Holmes Susan. Ten quick tips for effective dimensionality reduction. PLoS Comput Biol 2019;15(6):e1006907.

[30] Vogelstein Joshua T, Bridgeford Eric W, Tang Minh, Zheng Da, Douville Christopher, Burns Randal, et al. Supervised dimensionality reduction for big data. Nat Commun 2021;12(1):1–9.

[31] Jain Divya, Singh Vijendra. Feature selection and classification systems for chronic disease prediction: A review. Egypt Inform J 2018;19(3):179–89.