Comparative Analysis of Data Synthesis Methods for Prognostic Models Development in Cardiology

Vladimir V. Kosterin[1][0000-0003-3747-7438], Karina I. Shakhgeldyan[1,2][0000-0002-4539-685X], Boris I. Geltser[1,2] [0000-0002-9250-557X] and Vladislav Yu. Rublev[1,2][0000-0001-7620-4454]

[1] Vladivostok State University, Institute of Information Technologies, 41. Gogol str., Vladivostok, 690014, Russian Federation
[2] Far Eastern Federal University, School of Medicine, 10. Ajax Bay, build. 25, Vladivostok, 690920, Russian Federation
kosterin_vv@protonmail.com

Abstract. In recent years, machine learning (ML) methods have been widely used in various subject areas. Depending on which area we interfere, the cost of collecting big data and preparing it for further analysis can vary greatly. Medicine is one of those costly areas for data collection, which is often represented by unbalanced classes, combined with their total number limitation. The model accuracy largely depends on the initial data amount, which was used for the model training. The final predictive value of the model could be worsening in case of datasets imbalance and insufficient volume of the minority class. Oversampling methods are used to solve this problem, with the leading role of the SMOTE algorithm and its varieties. Previous researches have shown that oversampling algorithms have varying efficiency degrees, depending on the applying ML method. This paper represents the results of hypothesis testing about possibility of synthesized data usage in term of predicting the atrial fibrillation development and in-hospital mortality for patients with coronary heart disease after coronary artery bypass grafting. For model's development the following ML methods were used: multivariate logistic regression, stochastic gradient boosting and random forest. Data generation was performed by several varieties of the SMOTE algorithm. The analysis has shown that their usage for dataset extension in order to predict fatal and non-fatal cardiovascular events does not guarantee forecast quality improvement and in most cases leads to retraining of the models.

Keywords: Synthetic data, Oversampling methods, Machine learning, Artificial intelligence, Unbalanced sampling.

1 Introduction

Machine learning (ML) becoming an increasingly popular approach for the medical services development solving the problems of prediction in areas of disease development, their complications, and treatment outcomes [18]. The effective ML models development is associated with increased access to a large amount of medical data. Withal, one of the key problems of their application for

predicting the development of fatal and non-fatal events is the classes imbalance of the forecast end point and the insufficient amount of minority class data [2,8,12,16]. This situation is notably often recorded during prediction of surgical operations lethal outcomes, postoperative complications and development of other adverse events. Fatal events in cardiac surgery are happen relatively rarely (1-10%), which leads to a significant dataset's imbalance and insufficient prediction accuracy [11,13,14].

Various methods can be used to solve this class imbalance in medical research. One of them is balanced sampling, in which the number of records for each class becomes approximately the same. Several methods of balanced sampling have been developed, the use of which for the medical data analysis is constantly expanding [19,21]. At the same time, the question of this approach (when synthetic data is used to train predictive models) validity, remains insufficiently explored. Population researches are less sensitive to individual patients' characteristics, while the tasks of personalized medicine require taking into account the concrete clinical and functional state of patients.

This research aims to conduct a comparative effectiveness analysis of balanced sampling methods usage for predicting adverse events after cardiothoracic surgeries using ML methods.

2 Related works

In recent years, dozens of balanced sampling methods have been developed, which can be divided into 2 groups: data reduction (undersampling) and data augmentation (oversampling). The data reduction approach for a minority class can be applied in cases where the resulting dataset contains enough data for analysis [1,15]. Data augmentation methods based on objects of a minority class are best suited to a situation where the latter has a small volume [22]. An increase in the number of samples of a smaller class allows you to save all the information and develop a more balanced model, but the risks of its overfitting require additional researches on the validity of data synthesis [22]. In prediction problems in clinical medicine, where a small minority class size is most often observed, it is attractive to synthesize data using several methods. The basic method includes SMOTE, which generates a new synthetic example, placing it in the feature space between each example and its k-nearest neighbors in a minority class [6]. The working principle of Polynom-fit-SMOTE is to use polynomial curves to generate new synthetic examples that more accurately model the distribution of data in a minority class [9]. The CURE-SMOTE algorithm combines CURE clustering and SMOTE methods. First, the data is clustered using the CURE algorithm, and then the SMOTE algorithm is applied to generate synthetic samples for the minority class in each cluster [17]. To generate synthetic SOMO samples by calculating the neighborhood size for each sample and generating synthetic analogs by random perturbation inside their neighborhoods [7]. ProWSyn generates new synthetic samples using various techniques: copying and modifying existing objects, finding boundary examples and modifying them, and noise reduction using bagging [3]. The LoRAS algorithm is designed to generate synthetic analogues by approximating the main data set [4], while the ProWRAS algorithm integrates LoRAS and ProWSyn [5].

3 Methods and materials

This paper analyzes the effectiveness of data augmentation methods (oversampling) usage for prediction of the adverse events development for patients with coronary heart disease (CHD) after coronary bypass surgery (CABG). The research was performed on the dataset "Prognostic assessment of the clinical and functional status of patients with coronary artery disease after CABG"[*], which includes information on 999 patients who underwent in GBUZ "Primorsky Regional Clinical Hospital No. 1" in Vladivostok, a planned isolated CABG was performed. Two tasks were considered: the development of postoperative atrial fibrillation and the prediction of in-hospital mortality (IHM) as a complication of CABG. To solve the first task, 2 groups of individuals were identified, the first of which included 173 (19.5%) patients with newly diagnosed atrial fibrillation in the postoperative period, the 2nd - 716 (80.5%) patients without this complication. 110 patients with preoperative atrial fibrillation were excluded from the dataset. To solve the second task, 2 groups of people were identified among the examined cohort. The 1st of them included 63 (6.3%) patients who died in the hospital during the first 30 days after CABG (IHM), the 2nd included 936 (93.7%) patients with a favorable outcome of the operation. The selection and validation of predictors, as well as the prediction of endpoints, were previously performed by the authors of this work [10,20].

Based on the researches analysis results, the following oversampling methods were selected: SMOTE, Polynom-fit-SMOTE, ProWRAS, CURE-SMOTE, SOMO and ProWSyn, which were described as most effective [3,5]. Each of these methods has been shown to be best suited for certain ML models [5]. To predict IHM and the development of atrial fibrillation, ML methods were used: multivariate logistic regression, random forest, and stochastic gradient boosting. Cross-validation of the models was performed using the stratified K-Fold method for 10 samples. To assess the quality of the models, the following metrics were used: area under the ROC curve (AUC), sensitivity (Sen) and specificity (Spec), which were evaluated by averaging over 10 validating samples. In order to test the hypothesis about the possibility of using synthetic data for training models, the value of correctly predicted objects of class 1 (PPV) was estimated.

The research design included several stages (Fig. 1). For both prognostic tasks in the analyzed dataset, pools of potential predictors and dichotomous endpoints were determined. All predictors were additionally tested for the significance of differences in the comparison groups. For further research, only those of them were used that confirmed their significance as predictors of predicted events. From the dataset, based on a random stratified sample, 30% of objects were selected that had an end point equal to 1 (patients who died within 30 days after surgery - IHM and patients with atrial fibrillation). Data from this cohort of patients were not involved in sample synthesis, training, or cross-validation of models. They were used only for the final testing of models trained and validated on a combination of real and synthesized data by 6 oversampling methods. The remaining data (70% of objects with an endpoint of 1 and 100% of objects with an endpoint of 0) were used for training and cross-validation.

[*] Rublev V. Yu., Geltser B. I., Shakhgeldyan K. I. FEFU. State Registration Certificate No. 2022621907, publ. 08/02/2022, bul. #8

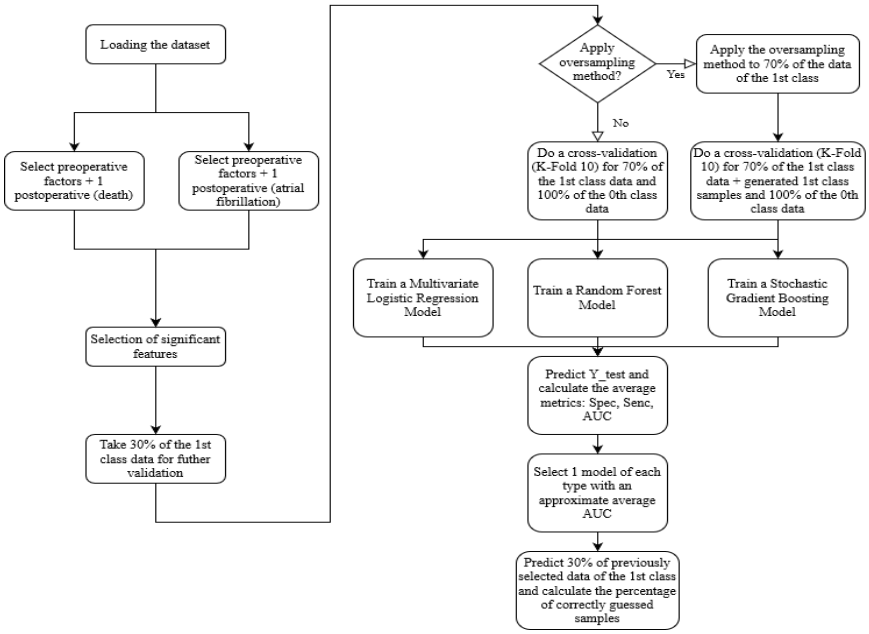

Fig.1. Research Design.

At the next stage, oversampling methods were applied to 70% of patients with an endpoint of 1 to obtain balanced datasets that included real and synthetic data generated based on them. This made it possible to form 6 combined datasets for 2 prognostic tasks. Each dataset was used for training and cross-validation using the stratified K-Fold method of three ML models. All models were trained with parameter and hyperparameter fitting to maximize AUC.

At the final stage of the research, all models, including those that were developed only on real data and those that were trained using combined samples, were tested on 30% of the previously separated data with an end point equal to 1. To do this, all models were trained on datasets of real and combined data. To evaluate the forecast results, we used the PPV metric, which is calculated as the ratio of predicted units to their total number 1 in the testing sample.

4 Results

The research considered 2 tasks on the same dataset, which differ in the size of the minority class. In the first task (forecasting atrial fibrillation), it was about 20%, in the second (forecasting IHM) - 6% of the data. To predict atrial fibrillation, a dataset of real clinical data was used, including indicators of 121 patients from class 1, 716 patients from class 0, and 595 “synthetic patients”, generated from 121 “real” patients. For the final testing, the indicators of 52 patients from class 1 were used, which were isolated from the real dataset before data synthesis. The most significant predictors of the Mann-Whitney test and the Chi-Square test for continuous and categorical variables, respectively, were selected as input data. The most significant predictors included: the age of

patients, the concentration of glucose in the blood, electrocardiogram parameters (PQ, QRS, QT), the size of the right atrium and the presence of chronic heart failure III-IV functional class. The results of training, cross-validation and final testing are shown in several tables (Table 1–3).

Table 1. Multivariate logistic regression quality metrics for predicting atrial fibrillation.

Oversampling method	Sensitivity	Specificity	AUC	PPV of 30% (class 1)
Without oversampling	0.66	0.64	0.68	53%
SMOTE	0.67	0.63	0.7	58%
ProWSyn	0.65	0.66	0.7	56%
SOMO	0.66	0.64	0.68	53%
CURE-SMOTE	0.91	0.81	0.91	53%
Polynom-fit-SMOTE	0.88	0.81	0.91	49%
ProWRAS	0.94	0.86	0.93	48%

Averaged quality metrics for cross-validation of prognostic models of atrial fibrillation after CABG based on multivariate logistic regression showed a significant advantage of the ProWRAS data synthesis method compared to using only real data (AUC=0.93 vs. 0.68, respectively). At the same time, the forecast quality of the model trained on synthesized data was lower than that of the model trained on real data (PPV=53% vs 48%). A slight increase in the forecast accuracy of the model trained on the combined data was demonstrated by the model using the data synthesized by the SMOTE and ProWSyn methods during training (PPV=58%, 56%).

Table 2. Random Forest model quality metrics for atrial fibrillation prediction.

Oversampling method	Sensitivity	Specificity	AUC	PPV of 30% (class 1)
Without oversampling	0.61	0.62	0.67	53%
SMOTE	0.8	0.79	0.86	31%
ProWSyn	0.79	0.82	0.87	27%
SOMO	0.65	0.61	0.67	56%
CURE-SMOTE	0.93	0.97	0.98	19%
Polynom-fit-SMOTE	0.94	0.91	0.98	21%
ProWRAS	0.94	0.93	0.97	29%

The random forest model for predicting atrial fibrillation on real data provided a low AUC (0.67) according to the results of cross-validation (see Table 2). At the same time, a significant increase in accuracy up to AUC=0.98 was observed on the combined sample when using the Polynom-fit-SMOTE and CURE-SMOTE synthesis methods. Testing models on real data showed a low generalizing ability of random forest models trained on combined data (PPV=21% and 19% versus 53% for the real dataset). The SOMO method provided a slight improvement in the quality of the forecast (PPV=56%), while maintaining other quality metrics (AUC=0.67). The remaining 5 synthesis methods led to significant overfitting and reduced the generalizing ability of predictive models.

Table 3. Stochastic Gradient Boosting quality metrics for atrial fibrillation prediction.

Oversampling method	Sensitivity	Specificity	AUC	PPV of 30% (class 1)
Without oversampling	0.61	0.62	0.67	53%
SMOTE	0.8	0.79	0.86	31%
ProWSyn	0.79	0.82	0.87	27%
SOMO	0.67	0.6	0.72	58%
CURE-SMOTE	0.92	0.89	0.97	9%
Polynom-fit-SMOTE	0.92	0.9	0.97	17%
ProWRAS	0.91	0.9	0.96	34%

Stochastic gradient boosting showed similar quality metrics and generalizing ability to random forest. Training and cross-validation had the best quality scores when using the Polynom-fit-SMOTE data synthesis method (AUC=0.97). At the same time, the final testing of 5 out of 6 synthesis methods demonstrated overfitting of the models (PPV=9%-34%). The model trained on real data provided the correctness of the estimate at the level of 53%. Only the SOMO method (PPV=58%) showed the best results with identical quality metrics on cross-validation (see Table 3).

To predict IHM, the same dataset was used, including indicators of 58 patients from class 1, 561 patients from class 0, and 503 "synthetic patients" generated from 58 "real patients". For the final testing, data from 17 patients from class 1 were used, which were isolated from the real dataset before data synthesis. The most significant predictors were selected, which included: age of patients, ejection fraction of blood from the left ventricle, end diastolic and systolic volume of the left ventricle, pulmonary artery pressure, sizes of the left and right atrium, blood parameters: hemoglobin, leukocytes, total protein, urea, prothrombin index, thrombin time, creatinine clearance, neutrophils; patients body weight, the duration of the QRS interval on the ECG, the presence of heart failure and angina III or IV functional classes, chronic kidney disease, recent myocardial infarction and extracardiac arteriopathy. The results of training, cross-validation and final testing are shown in the following tables (Table 4-6).

Table 4. Multivariate Logistic Regression quality metrics for IHM prediction.

Oversampling method	Sensitivity	Specificity	AUC	PPV of 30% (class 1)
Without oversampling	0.6	0.75	0.7	76%
SMOTE	0.83	0.81	0.91	58%
ProWSyn	0.86	0.8	0.9	58%
SOMO	0.6	0.75	0.69	76%
CURE-SMOTE	0.91	0.81	0.91	64%
Polynom-fit-SMOTE	0.88	0.81	0.91	58%
ProWRAS	0.9	0.86	0.93	58%

The analysis showed that the use of synthetic data for training a multivariate logistic regression model did not lead to an improvement in the quality of the forecast with any oversampling method. Cross-validation quality metrics were

identical, and generalizability indicators showed performance degradation across all methods except SOMO, where they were comparable to baseline results (PPV=76%).

Table 5. Random Forest quality metrics for IHM prediction.

Oversampling method	Sensitivity	Specificity	AUC	PPV of 30% (class 1)
Without oversampling	0.75	0.73	0.82	58%
SMOTE	0.9	0.89	0.96	47%
ProWSyn	0.88	0.89	0.96	41%
SOMO	0.75	0.74	0.82	58%
CURE-SMOTE	0.93	0.97	0.98	5%
Polynom-fit-SMOTE	0.94	0.91	0.98	11%
ProWRAS	0.94	0.93	0.97	35%

Random forest models showed similar multivariate logistic regression sensitivity to the use of synthetic data with a significantly (<10%) imbalanced sample (see Table 5). The best results of random forest models were obtained using only real data and a combined dataset synthesized by the SOMO method with data (PPV=58%).

Table 6. Stochastic Gradient Boosting quality metrics for IHM prediction.

Oversampling method	Sensitivity	Specificity	AUC	PPV of 30% (class 1)
Without oversampling	0.68	0.71	0.77	64%
SMOTE	0.9	0.91	0.96	41%
ProWSyn	0.89	0.9	0.96	29%
SOMO	0.68	0.71	0.77	64%
CURE-SMOTE	0.92	0.89	0.97	41%
Polynom-fit-SMOTE	0.92	0.9	0.97	29%
ProWRAS	0.91	0.9	0.96	41%

The stochastic boosting method provided better generalization abilities compared to random forest, but the quality metrics based on the results of cross-validation were identical. The best generalization ability was recorded using real data or those synthesized by the SOMO method (PPV=64%).

5 Discussion

The present research was devoted to the effectiveness analysis of synthetic data usage for predictive models development in clinical medicine on unbalanced samples. The design of the research ensured the correctness of hypothesis testing about the ability of models trained on synthetic (combined) data to predict adverse cardiac surgery events. The authors separated 30% of the real data of the minority class from the processes of data synthesis, training and cross-validation of models. In this paper, we analyzed the possibilities of 6 data synthesis methods and 3 ML methods for developing predictive models, which were used to solve 2

binary classification problems with different ratios of the minority and majority classes.

Previous researches have shown that each of the oversampling methods is best suited for certain ML models, which is weakly consistent with the results of our research [5]. We have not found a significant increase in the quality of the forecast with any oversampling method. Moreover, with a significant amount of synthetic data compared to real ones (the ratio is 10), most of the oversampling methods, except for SOMO, showed a decrease in the generalizing abilities of ML models. The SOMO model did not improve the forecast quality. With smaller ratios of the amount of synthetic and real data (a ratio of around 3), there was a slight improvement in the generalizing ability of the multivariate logistic regression model built on data synthesized by the SMOTE method. Random Forest and Stochastic Gradient Boosting models had similar results on SOMO-generated data.

It is also important to note that the cross-validation results obtained from combined samples synthesized by all methods except SOMO cannot be considered correct, as they showed clear signs of overfitting. The latter were showing greater results, depending of the ratio of the synthesized data to real one's volume. The research is limited by clinical data and the methods considered, and should not be translated into images or time series and other data synthesis methods.

6 Conclusion

The research results indicate that there is no effect of increasing the accuracy of ML models from the use of 6 data synthesis methods in the clinical medicine prognostic tasks, particularly, in cardiac surgery. This indicates the requirement of new synthesis methods development, the use of which will increase the predictive properties of ML models.

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

Acknowledgments

The study was supported by the Russian Science Foundation grant No. 23-21-00250, https://rscf.ru/project/23-21-00250/

References

1. Alam, T., Shaukat, K., Hameed, I., et al: A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining. Biomedical Signal Processing and Control, 68 (2021).
2. Aseeri, M., Hassanat, A., Mnasri, S.: Modelling-based simulator for forecasting the spread of COVID-19: A case study of Saudi Arabia. International Journal of Computer Science and Network Security, 20, 114-125. (2020).

3. Barua, S., Islam, M., Murase, K.: Advances in Knowledge Discovery and Data Mining. ProWSyn: Proximity Weighted Synthetic Oversampling Technique for Imbalanced Data Set Learning, pp. 317-328. Heidelberg: Springer-Verlag (2013).
4. Bej, S., Davtyan, N., Wolfien, M., et al: LoRAS: An oversampling approach for imbalanced datasets. Machine Learning, 110(2), 279-301. (2021).
5. Bej, S., Schulz, K., Srivastava, P., et al: A Multi-Schematic Classifier-Independent Oversampling Approach for Imbalanced Datasets. IEEE Access, 9, 123358-123374. (2021).
6. Chawla, N., Bowyer, K., Hall, L., et al: SMOTE: Synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. (2002).
7. Douzas, G., Bacao, F.: Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. Expert Systems with Applications, 82, 40-52. (2017).
8. Fatima, M., Pasha, M.: Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(1), 1-16. (2017).
9. Gazzah, S., Essoukri, N.: The 8th IAPR Workshop on Document Analysis. New Oversampling Approaches Based on Polynomial Fitting for Imbalanced Data Sets (pp. 677-684). Nara: DAS. (2008)
10. Guo, X., Yin, Y., Dong, C., et al: Proceedings of the 4th International Conference on Natural Computation. On the class imbalance problem (pp. 192-201). Jinan: IEEE. (2008).
11. Hammad, M., Alkinani, M., et al: Myocardial infarction detection based on deep neural network on imbalanced data. Multimedia Systems, 1-13. (2021).
12. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York: USA: Springer. (2009).
13. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9, 1263-1284. (2009).
14. Jindaluang, W., Chouvatut, V., Kantabutra, S.: Proceedings of the International Computer Science and Engineering Conference. Under-sampling by algorithm with performance guaranteed for class-imbalance problem. (pp. 215-221). Khon Kaen: ICSEC. (2014).
15. Li, D., Liu, C., Hu, S.: A learning method for the class imbalance problem with medical data sets. Computers in Biology and Medicine 40(5), 509-518. (2010).
16. Ma, L., Fan, S.: CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinformatics,18(1), 169. (2017).
17. May, M: Eight ways machine learning is assisting medicine. Nature Medicine 27, 2–3. (2021).
18. Ramezankhani, A., Pournik, O., Shahrabi, J., et al: The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. Medical Decision Making, 36(1), 137–143. (2014).
19. Shakhgeldyan, K., Rublev, V., Geltser, B., Shcheglov, B., Shcheglova, S.: Metody mashinnogo obucheniya dlya prognozirovaniya riska razvitiya fibrillyatsii predserdiy posle koronarnogo shuntirovaniya. Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte (IMMV-2021). vol. 2, pp. 269-283. Smolensk: Universum. (2021).
20. Shakhgeldyan, K., Rublev, V., Geltser, B., Shcheglov, B., Kosterin, V., Shcheglova, S.: Metody mashinnogo obucheniya dlya prognozirovaniya riska vnutrigospital'noy letal'nosti posle koronarnogo shuntirovaniya. Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte (IMMV-2022), Volume 1, pp. 274-286. Kolomna: Russian Association for Artificial Intelligence. (2022).

21. Turlapati, V. P., Prusty, M.R.: Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. Intelligence-Based Medicine, 3-4. (2020).
22. Zheng, Z., Cai, Y., Li, Y.: Oversampling method for imbalanced classification. Computers & Informatics, 34(5), 1017-1037. (2015).