INTEGRATION OF SVM AND SMOTE-NC FOR CLASSIFICATION OF HEART FAILURE PATIENTS

Dina Tri Utari[1*]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Indonesia Islamic University
Kaliurang Street Km 14.5, Sleman, Yogyakarta, 55584, Indonesia

Corresponding author's e-mail: * dina.t.utari@uii.ac.id

ABSTRACT

SMOTE (Synthetic Minority Over-sampling Technique) and SMOTE-NC (SMOTE for Nominal and Continuous features) are variations of the original SMOTE algorithm designed to handle imbalanced datasets with continuous and nominal features. The primary difference lies in their ability to generate synthetic examples for the minority class when dealing with continuous and nominal features. We employed a dataset comprising continuous and nominal features from heart failure patients. The distribution of patients' statuses, either deceased or alive, exhibited an imbalance. To address this, we executed a data balancing procedure using SMOTE-NC before conducting the classification analysis with SVM. It was found that the combination of SVM and SMOTE-NC methods gave better results than the SVM method, seen from the higher level of accuracy and F1 score. F1 gives less sensitivity to class imbalance compared to accuracy. Suppose there is a significant imbalance in the number of instances between classes. In that case, the F1 score can be a more informative metric for evaluating a classifier's performance, especially when the minority class is of interest.

1. INTRODUCTION

Along with technology development, the amount of information available on the internet has been distributed globally and reasonably significantly. Nevertheless, if this information is not used, it will only become a useless data collection. This information allows people to find and interpret patterns that can help make decisions, one of which is data mining techniques. One of the data mining techniques used to predict a decision is classification [1]. The classification has several algorithms based on fuzzy logic, Bayes classification, decision tree, support vector machine, artificial neural network, and k-nearest neighbor.

Support Vector Machine (SVM) has better accuracy than the k-nearest neighbor, decision tree, and linear regression algorithms [2]. The experimental findings indicate that the SVM classifier achieves the highest accuracy and demonstrates the most heightened sensitivity following the training and testing conducted using the proposed method. SVM is used to find the optimal hyperplane that separates the two classes to provide good generalization capabilities. We usually take most of the label data to find the optimal hyperplane. However, large-scale experimental data can create complexity the higher the computational process [3].

According to World Health Organization, an estimated 17.9 million people died from heart failure in 2019, representing 32% of all [4]. Many factors cause a person to have heart failure, such as age, high blood pressure, and other factors. We used heart failure datasets from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) with 13 attributes such as age of patient, the diabetes status, and more.

Based on heart failure dataset, it was found that only a tiny proportion of patients died compared to those who survived. This condition indicates that the death event is imbalanced between the dead (minority) and survived (majority) classes. Imbalanced datasets are those in which the number of instances in one class is much greater than the number of examples in another [5]. Insights from such datasets can influence decision results. Hence the practical ramifications of uneven training data are considerable. Unbalanced data distributions are widespread in real-world contexts, especially when target classes lack uniform distribution across various tiers [6]. This problem is prevalent in healthcare applications, where a class imbalance is a significant obstacle [6].

The most popular method of dealing with data imbalances is resampling. One of the resampling methods that is quite popular is the Synthetic Minority Oversampling Technique (SMOTE). The SMOTE method works by replicating data from minority data based on the k-nearest neighbors to be equivalent to the majority class [7]. Of those popular resampling methods, the author applies them to the case of this research data, which has an imbalanced class. However, because the data used are nominal and continuous, this study proposes a class imbalance handling method using the SMOTE-NC algorithm proposed by Gök and Olgun [8].

This study will apply the integration of the oversampling SMOTE-NC method with the SVM classification in predicting the death event of a patient's heart failure, whether the object is in the dead or survived class. Imbalanced datasets that will be analyzed by SVM can introduce bias during learning, as the model may focus more on the prevalent class. Balancing mitigates this bias, allowing the model to learn the underlying patterns in each class more effectively. In medical diagnosis applications, the varying costs associated with false positives and false negatives stemming from data imbalance underscore the need to integrate classification algorithms seamlessly with effective data balancing techniques. This integration is essential for achieving superior and more reliable results in the diagnostic process.

2. RESEARCH METHODS

2.1 Support Vector Machine

Classification is a process of finding and determining a model or function that can explain and differentiate data classes to use the data to estimate the class of an object whose status is unknown. During the learning process in modeling, a learning algorithm is needed, including Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (K-NN), Decision Tree, Artificial Neural Network (ANN), and others. SVM is one of the classification methods in data mining. SVM can also predict both classification and regression. SVM was first proposed by Vladimir Vapnik and is used in statistical learning theory and

structural risk minimization [9]. SVM are commonly employed in healthcare challenges like disease prediction [10]. SVM has been enhanced in survival analysis to handle censored data, making it suitable for predicting survival and discharge-time likelihood in COVID-19 cases [10]. Various machine learning algorithms, including SVM, are utilized in the healthcare business to diagnose disorders [11]. SVM is utilized as a classifier in a diagnosis model for outpatient clinicians employing healthcare data analytics and neural networks [12]. SVM is also utilized in a clinical decision support system to predict heart illness [13]. The research addresses the computational difficulties of solving nonlinear SVM on large-scale data with uneven class sizes [14].

SVM has a linear principle, but now SVM has developed to work on non-linear problems. The way SVM works on non-linear issues is to incorporate the kernel concept into a high-dimensional space. In this dimensional space, a separator or what is often called a hyperplane will be searched. Hyperplane can maximize the distance or margin between data classes. The best hyperplane between the two categories can be found by measuring the margin and finding the maximum point. The effort to find the best hyperplane in the class separator is the core of the process in the SVM method.

Consider the n-dimensional training dataset $T = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ where m denotes the number of samples, x_i is a sample in the input space X, and $y_i \in \{-1, 1\}$ is the label in the output Y. The support vector machine's core principle is to discover an ideal hyper-plane that maximizes the area of both sides of the hyper-plane. The separation hyper-plane for standard linear se accuracy of classification a support vector classification (SVC) can be defined as [15]:

$$f(x) = w^T x + b = 0 \tag{1}$$

where w is a vector weight, $w = \{w_1, w_2, \ldots, w_n\}$, n is the number of attributes, and b is a scalar, often called bias. If two classes are not completely separable, a hyperplane can still be determined to maximize the margin while minimizing a quantity proportional to the misclassification errors with add positive slack variables ξ, so that the optimization problem can be described as follows:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \xi_i \tag{2}$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

where C is a parameter chosen by user that controls the tradeoff between margin and the misclassification errors.

The challenge of maximizing the interval is turned into a convex quadratic programming problem using the dual approach in convex optimization. SVMs use nonlinear kernels to translate low-dimensional feature space to high-dimensional space to deal with nonlinear challenges, the kernel function is defined as: $K(x_i, x_j) = \langle \phi(x_i). \phi(x_j) \rangle$, and $\phi(x_i)$ is a mapping [16]. **Table 1.** shown some typical kernels function [16]. The kernel trick enables SVM to perform computations in the original feature space without explicitly mapping data into the higher-dimensional space.

Table 1. Kernels Functions

Type of Kernel	Kernel Function
Linear	$K(x_i, x_j) = (x_i^T x_j)$
Polynomial	$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
Radial Basis Function (RBF)	$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right), , \gamma > 0$

γ, r, and d are kernel parameters.

2.2 Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC)

For nominal and continuous mixed datasets, the method proposed by Chawla is called Synthetic Minority Oversampling Technique Nominal Continuous (SMOTE-NC). The SMOTE-NC algorithm is described below [8].

1. Calculate Median: Calculates the median of the standard deviations of all consecutive characteristics of the minority class. If the nominal attributes differ between the sample and its potential nearest neighbors, this median is included in the Euclidean distance calculation. Use the median to penalize nominal feature differences in quantities related to general differences in continuous feature values.

2. Nearest Neighbor Calculation: Uses continuous feature space to calculate the Euclidean distance between a feature vector (minority class sample) that identifies the nearest neighbor and another feature vector (minority class sample). Each nominal feature deviates between the feature vector considered and its nearest potential, including the median of the previously calculated standard deviation in the Euclidean distance calculation.

3. Populate the artificial pattern: The non-stop capabilities of the brand-new artificial minority class pattern are created using the equal SMOTE method, as defined earlier. The nominal characteristic is given the price occurring inside the majority of the k-nearest neighbors.

3. RESULTS AND DISCUSSION

3.2 Materials

The data used in the study was data of the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan), during 13 attributes [17]. This study aimed to classify death events: the survived patients (death event = 0) and the dead patients (death event = 1). Before the classification, handling imbalanced data between classes on the death event variable using SMOTE-NC is carried out.

Table 2. Attributes Description of the Hearth Failure Patients Dataset

| No	Attributes	Description	Type
1	Age	Age of patient	Continuous
2	Anemia	Decrease of red blood cells or hemoglobin	Nominal
3	Creatinine phosphokinase	Level of the CPK enzyme in the blood	Continuous
4	Diabetes	If the patient has diabetes	Nominal
5	Ejection fraction	Percentage of blood leaving of blood leaving	Continuous
6	High blood pressure	If a patient has hypertension	Nominal
7	Platelets	Platelets in the blood	Continuous
8	Serum creatinine	Level of creatinine in the blood	Continuous
9	Serum sodium	Level of sodium in the blood	Continuous
10	Sex	Woman or man	Nominal
11	Smoking	If the patient smokes	Nominal
12	Time	Follow-up period	Continuous
13	Death Event	If the patient died during the follow-up period	Nominal

3.2 Descriptive Statistics

In this study, we used 13 attributes with seven continuous attributes and six nominal type attributes. The aim of this study is to classify death events based on the attributes that influence them. The following bar plot shows the death event category, namely survived patient, and dead patient from a total of 299 data.
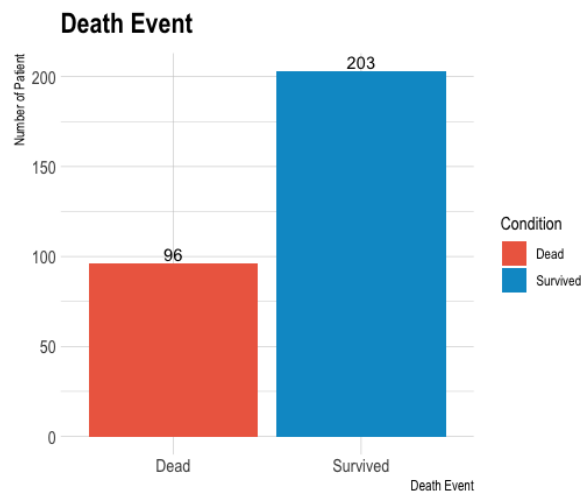
Figure 1. Total Death Events by Patient Status

Figure 1. explains that from a total of 299 data on heart failure patients, 203 patients were survived and 96 patients died. The next stage is the division of the dataset into training and testing data with a proportion of 75%:25%. In cases where there is an imbalance class, the accuracy of the classification results in the survived class (majority) will be high. In contrast, the classification accuracy in the dead class (minority) will tend to below. The classifier will consider minor classes, in this case is dead status, as noise or outliers when forming a classification function. In addition, the dataset used consists of a mixture of nominal and continuous attributes. Therefore, it is necessary to handle imbalanced data using SMOTE-NC. The following are the results after oversampling.
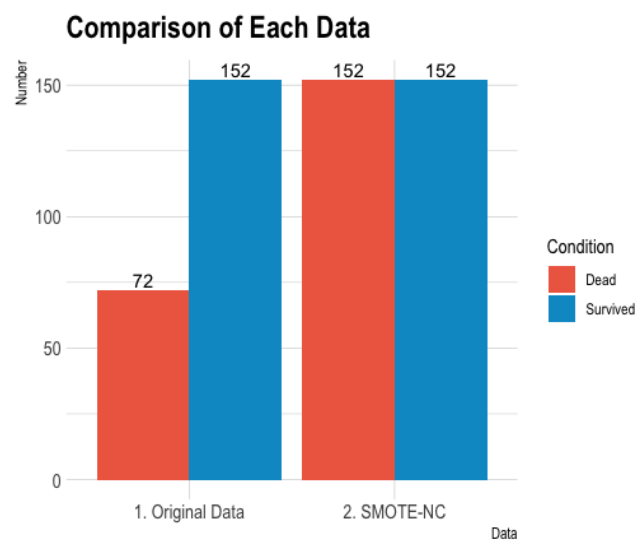
Figure 2. Training Data Balancing Results

Based on **Figure 2.**, after balancing the data using SMOTE-NC, the distribution of the dead (minority) class is balanced with the survived (majority) class. SMOTE-NC extends the concept of SMOTE to handle nominal features by generating synthetic samples not only in the continuous feature space but also in the nominal feature space.

For nominal features, SMOTE-NC randomly selects a feature value from the nearest neighbors to create synthetic instances, preserving the distribution of the original nominal feature values. By generating synthetic samples for the minority class, SMOTE-NC helps balance the class distribution in the dataset.

3.3 SVM Analysis

After splitting the dataset, the next step is classification using SVM. Training data which is 224 data is used to build the SVM model, while testing data of 75 for model evaluation. This study will compare the classification results to the original data and balanced data from SMOTE-NC results.

a) SVM for Original Data

In SVM modeling, the C or cost and gamma values are first determined. Next, in this study, we try 35 combinations for each parameter. The best C and gamma values have the lowest error rates. **Figure 3.** below is the comparison of 35 combination of C and gamma values. C, a hyperparameter in SVM, is a regularization parameter that influences the trade-off between achieving a smooth decision boundary and correctly classifying training points. A low C emphasizes a more generalized decision boundary, allowing for some misclassification but preventing overfitting. Conversely, a large C imposes a stricter requirement for correct classification, potentially leading to a decision boundary that closely fits the training data but may need to generalize better to new data.

On the other hand, gamma comes into play when employing the Gaussian Radial Basis Function (RBF) kernel in SVM. Setting gamma is a crucial step before training the model. This parameter determines the extent of curvature in the decision boundary. A more minor gamma results in a broader, more gradual decision boundary, while a larger gamma leads to a more intricate and tightly fitted decision boundary that can capture intricate patterns in the training data.

In essence, selecting appropriate values for C and gamma involves finding a balance that aligns with the complexity of the dataset.

Table 3. Combination of C and Gamma Values of SVM

Cost	Gamma	Error Rate	Cost	Gamma	Error Rate
0.001	0.01	0.321	10	1	0.330
0.01	0.01	0.321	100	1	0.330
0.1	0.01	0.321	1000	1	0.330
1	0.01	0.214	0.001	10	0.321
10	**0.01**	**0.170**	0.01	10	0.321
100	0.01	0.214	0.1	10	0.321
1000	0.01	0.214	1	10	0.321
0.001	0.1	0.321	10	10	0.321
0.01	0.1	0.321	100	10	0.321
0.1	0.1	0.321	1000	10	0.321
1	0.1	0.232	0.001	100	0.321
10	0.1	0.218	0.01	100	0.321
100	0.1	0.238	0.1	100	0.321
1000	0.1	0.238	1	100	0.321
0.001	1	0.321	10	100	0.321
0.01	1	0.321	100	100	0.321
0.1	1	0.321	1000	100	0.321
1	1	0.325			

Table 3. shows that the best parameters when tuning parameters are obtained: cost = 10, gamma = 0.01, and RBF kernel function. The RBF kernel allows non-linear decision boundaries to be represented in the original feature space by implicitly mapping the input data into a higher-dimensional space. This capacity is essential when working with intricate patterns that are difficult to represent fully in a lower-dimensional area. Furthermore, the model will be built on the training data using these parameters and make predictions on the testing data.

Table 4. Confusion Matrix of SVM Model

		Actual	
		Dead	**Survived**
Prediction	**Dead**	43	7
	Survived	8	17

Table 5. SVM Model Evaluation

Measurement	Value
Accuracy	80.00%
Sensitivity	68.00%
Specificity	86.00%
Presicion	70.83%
F1 score	69.39%
AUC	0.77

Table 4. and Table 5. shows that heart failure patients who fall into the dead category and are correctly predicted to be dead are 43. Likewise for the survived category, which correctly predicted as many as 17. The performance of the quality classification model has an accuracy of 80.00%, which means that the accuracy of the model in classifying dead and survived category is 80.00%. The sensitivity is 68.00%, which means that the model can correctly predict dead (minority class) of 68.00%. The specificity of 86.00% means that the model can correctly predict survived (majority class) of 86.00%. The precision is 70.83%, which means that the accuracy of the model in classifying dead is 70.83%. The F1 score is 69.39%, which means the classification model has a precision and a sensitivity of 69.39%. In addition to the F1 score, the AUC value can also be used to see the model's performance. In this study, an AUC value of 0.77 was obtained, meaning that the model can classify, and the assessment of dead and survived category can be said to be quite good.

b) SVM and SMOTE-NC

In principle, SMOTE is an oversampling algorithm that works to duplicate by creating synthetic data in the minor class based on the k-nearest neighbors. Since the dataset used consists of nominal and continuous data, this synthesis data is obtained by calculating the closest distance to the observations to be duplicated. The number of duplications carried out is determined so that the ratio between major and minor classes plus synthetic minor classes will approach balance. Furthermore, classification using SVM with SMOTE-NC uses the combination of C and gamma values same as previous method.

Table 6. Combination of C and Gamma Values of SVM and SMOTE-NC

Cost	Gamma	Error Rate	Cost	Gamma	Error Rate
0.001	0.01	0.546	10	1	0.319
0.01	0.01	0.546	100	1	0.319
0.1	0.01	0.300	1000	1	0.319
1	**0.01**	**0.165**	0.001	10	0.563
10	0.01	0.171	0.01	10	0.563
100	0.01	0.207	0.1	10	0.563
1000	0.01	0.211	1	10	0.529
0.001	0.1	0.543	10	10	0.523
0.01	0.1	0.543	100	10	0.523
0.1	0.1	0.267	1000	10	0.523
1	0.1	0.181	0.001	100	0.563
10	0.1	0.197	0.01	100	0.563
100	0.1	0.227	0.1	100	0.563

Cost	Gamma	Error Rate	Cost	Gamma	Error Rate
1000	0.1	0.227	1	100	0.549
0.001	1	0.556	10	100	0.549
0.01	1	0.556	100	100	0.549
0.1	1	0.556	1000	100	0.549
1	1	0.329			

Different from the SVM results in the original data, in the combination of SVM and SMOTE-NC, it was obtained that the best results for each parameter were cost = 1, gamma = 0.01, and RBF kernel function. Therefore, that the confusion matrix for this model is obtained as follows,

Table 7. Confusion Matrix of SVM and SMOTE-NC Model

		Actual	
		Dead	**Survived**
Prediction	**Dead**	40	8
	Survived	11	43

By using testing data as much as 102 with each condition of the patient dead or survived in balance, it was obtained that heart failure patients who fall into the dead category and are correctly predicted to be dead are 40. Likewise for the survived category, which correctly predicted as many as 43. The table below compares evaluations for the SVM model and the SVM model combined with SMOTE-NC.

Table 8. Comparison of Model Evaluation

Measurement	SVM	SVM and SMOTE-NC
Accuracy	80.00%	81.37%
Sensitivity	68.00%	79.63%
Specificity	86.00%	83.33%
Presicion	70.83%	84.31%
F1 score	69.39%	81.90%
AUC	0.77	0.81

Based on the comparison results of several model evaluation measures in **Table 8.**, the model with SVM and SMOTE-NC can be said to have better performance than the SVM model because it has a higher F1 score. This is because, in the case of imbalanced data, or the dataset that we have, the number of False Negative and False Positive data is not close; we use the F1 score as a reference.

4. CONCLUSIONS

a. Integrating the SMOTE-NC technique with SVM proved to be a robust approach for addressing imbalanced datasets containing continuous and nominal data. By effectively balancing the data distribution, SMOTE-NC paved the way for enhanced classification accuracy when utilizing SVM in the context of heart failure patient datasets.

b. The findings from our study demonstrated a notable improvement in classification performance, showcasing the synergistic impact of SMOTE-NC and SVM. Notably, the classification model achieved an impressive precision and sensitivity of 81.90%, underscoring its ability to discern between different outcomes in heart failure patients accurately. These results affirm the efficacy of SMOTE-NC as a valuable preprocessing step for imbalanced data and highlight the complementary strengths of SVM in handling the intricacies of heart failure patient datasets. The heightened F1 score further emphasizes the balanced precision and recall achieved by the model, reinforcing its reliability in capturing true positive instances while minimizing false positives and false negatives.

REFERENCES

[1]   S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," in *2013 International Conference on Machine Intelligence and Research Advancement*, 2013.

[2]   A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive Analytics for Chronic Kidney Disease using Machine Learning Technique," in *2016 Management and Innovation Technology International Conference (MITicon)*, 2016.

[3]   Y. Yao *et al.*, "K-SVM: An Effective SVM Algorithm Based on K-Means Clustering," *J Comput (Taipei)*, vol. 8, no. 10, pp. 2632–2639, 2013.

[4]   World Health Organization, "cardiovascular diseases (CVDs)," *World Health Organization*, Jun. 11, 2021.

[5]   L. Mathews and S. Hari, "Learning from Imbalanced Data," 2019, pp. 403–414. doi: 10.4018/978-1-5225-7598-6.ch030.

[6]   A. E. Karrar, "Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130122.

[7]   R. A. Barro and D. S. Itasia, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang pada Pembuatan Model Komposisi Jamu," *Xplore*, vol. 1, no. 1, pp. 1–6, 2013.

[8]   E. C. Gök and M. O. Olgun, "SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples," *Neural Comput Appl*, vol. 33, pp. 15693–15707, 2021.

[9]   J. Nayak, B. Naik, and H. S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges," *International Journal of Database Theory and Application*, vol. 8, no. 1, pp. 169–186, 2015.

[10]   M. Nemati, J. Ansary, and N. Nemati, "Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data," *Patterns*, vol. 1, no. 5, p. 100074, Aug. 2020, doi: 10.1016/j.patter.2020.100074.

[11]   I. Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, Mar. 2021, doi: 10.38094/jastt20179.

[12]   M. S. Hossain and G. Muhammad, "Healthcare Big Data Voice Pathology Assessment Framework," *IEEE Access*, vol. 4, pp. 7806–7815, 2016, doi: 10.1109/ACCESS.2016.2626316.

[13]   N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," *IEEE Access*, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[14]   E. Sadrfaridpour, T. Razzaghi, and I. Safro, "Engineering fast multilevel support vector machines," *Mach Learn*, vol. 108, no. 11, pp. 1879–1917, Nov. 2019, doi: 10.1007/s10994-019-05800-7.

[15]   L. Cao and H. Shen, "Imbalanced data classification based on hybrid resampling and twin support vector machine," *Computer Science and Information Systems*, vol. 14, no. 3, pp. 579–595, 2017, doi: 10.2298/CSIS161221017L.

[16]   H. Xi and T. Chang, "Image Classification Based on Histogram Intersection Kernel," *Journal of Computer and Communications*, vol. 03, no. 11, pp. 158–163, 2015, doi: 10.4236/jcc.2015.311025.

[17]   D. Chicco and G. Jurman, "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone," *BMC Med Inform Decis Mak*, vol. 20, no. 16, pp. 1–16, 2020.