**Recursive Journal of Informatics**
Vol. 2, No. 2, September 2024
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

# Implementation of Random Forest with Synthetic Minority Oversampling Technique and Particle Swarm Optimization for Predicting Survival of Heart Failure Patients

**Untsa Zaaidatunni'mah[1], Endang Sugiharti[2]**
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

**Abstract.** Heart failure is caused by a disruption in the heart's muscle wall, which results in the heart's inability to pump blood in sufficient quantities to meet the body's demand for blood. The increasing prevalence and mortality rates of heart failure can be reduced through early disease detection using data mining processes. Data mining is believed to aid in discovering and interpreting specific patterns in decision-making based on processed information. Data mining has also been applied in various fields, one of which is the healthcare sector. One of the data mining techniques used to predict a decision is the classification technique.
**Purpose:** This research aims to apply SMOTE and PSO to the Random Forest classification algorithm in predicting the survival of heart failure patients and to determine its accuracy results.
**Methods/Study design/approach:** To predict the survival of heart failure patients, we utilize the Random Forest classification algorithm and incorporate data imbalance handling with SMOTE and feature selection techniques with PSO on the Heart Failure Clinical Records Dataset. The data mining process consists of three distinct phases.
**Result/Findings:** The application of SMOTE and PSO on the Heart Failure Clinical Records Dataset in the Random Forest classification process resulted in an accuracy rate of 93.9%. In contrast, the Random Forest classification process without SMOTE and PSO resulted in an accuracy rate of only 88.33%. This indicates that the proposed method combination can optimize the performance of the classification algorithm, achieving a higher accuracy compared to previous research.
**Novelty/Originality/Value:** Data imbalance and irrelevant features in the Heart Failure Clinical Records Dataset significantly impact the classification process. Therefore, this research utilizes SMOTE as a data balancing method and PSO as a feature selection technique in the Heart Failure Clinical Records Dataset before the classification process of the Random Forest algorithm.

**Keywords**: Data Mining, Heart Failure, Random Forest, SMOTE, PSO
**Received** November 08, 2023 / **Revised** May 02, 2024 / **Accepted** September 17, 2024



## INTRODUCTION

Data mining, also known as data extraction, is a relatively fast and straightforward technique for automatically discovering knowledge, patterns, and/or relationships within data [1]. As stated by [2] in their book, data mining is more accurately described as the process of extracting valuable knowledge from a set of raw data. Therefore, data mining is believed to assist in identifying and interpreting specific patterns when making decisions based on processed information.

In the field of healthcare, many researchers have utilized data mining as one of the methods to predict diseases, one of which is heart failure. This is because decision-making based on accurate data and information is likely to result in precise disease predictions and targeted interventions [3].

Heart failure is a complex set of symptoms caused by disruptions in the functioning of the heart [4]. The initial cause of heart failure is a disruption in the heart muscle's walls. Weakened heart muscle walls result in the heart's inability to pump and supply the body with the necessary blood and oxygen [5].

[1]*Corresponding author.
Email addresses: untsazaa@students.unnes.ac.id (Zaaidatunni'mah)
DOI: 10.15294/rji.v2i2.76142

Based on data from the Global Health Data Exchange (GHDx) in 2020, the worldwide number of heart failure cases reached 64.34 million, resulting in 9.91 million deaths [6]. The high prevalence of heart failure cases globally has spurred researchers to conduct studies aimed at early detection of this condition. According to [7], early detection, control, and management of the disease are crucial for preventive measures. One such study on heart failure is the research conducted by [8], which focuses on predicting the survival of heart failure patients using Genetic Algorithms and Adaptive Neuro Fuzzy Inference System based on the Heart Failure Clinical Records Dataset.

The utilization of data mining to predict a decision undoubtedly involves data classification techniques. Classification is a fundamental form of data analysis [9]. It falls under supervised learning, which can be used to categorize data or predict future data trends [10]. The classification process consists of two phases. The first phase is the learning process, where the training data is studied or analyzed by the classification algorithm to generate a model or classifier presented in the form of patterns or classification rules. The second phase involves using the model for classification and testing it on testing data to estimate the accuracy of the classification rules. One of the data mining classification methods is the Random Forest classification method.

Random Forest is an ensemble learning method that generates multiple decision trees as base classifiers, which are built and combined, and then performs majority voting to merge the results from each of these decision trees [11]. Random Forest has been recognized as a powerful ensemble classification method and works well in data processing. It has also been widely applied to various classification and regression tasks, which are also types of ensemble learning [12]. One study that utilizes Random Forest is the research conducted by [13] to predict the probability of heart failure using Random Forest. The results of their research produced a simple yet high-performing classification model, achieving a relatively good accuracy rate of 82.6087%. However, the Random Forest algorithm does not take into account data imbalance, which is often a common issue in the classification process of large datasets [14].

The issue of data imbalance can be addressed by using oversampling methods. Oversampling is a technique that generates new data or objects in the minority class, thereby balancing the minority and majority classes [15]. One commonly used oversampling method is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE aims to balance the class distribution by increasing the number of minority class data through the creation of synthetic data [16]. In the generation of synthetic data, SMOTE works by randomly selecting samples from the minority class and then finding their nearest neighbors among the chosen samples [17].

In addition to data imbalance, the classification process can also be disrupted when dealing with irrelevant features in the data. Excessive and irrelevant features can reduce classification performance and lower accuracy levels [18]. Therefore, a feature selection technique is needed to select relevant features. Feature selection is a step in simplifying a dataset by reducing dimensions and identifying relevant features without compromising prediction accuracy [19]. The feature selection method used in this research is Particle Swarm Optimization (PSO).

PSO is a metaheuristic algorithm proposed by J. Kennedy and R.C. Eberhart, which simulates social behavior, like a flock of birds flying to find the best positions to reach a specific goal in multidimensional space [20]. PSO is known for its superior search performance in solving optimization problems, offering faster convergence rates and stability [21]. Furthermore, the algorithm's simplicity and strong performance have captured the attention of researchers, leading to its application in various optimization problems [22].

Based on the description of the problem above, this research is focused on predicting the survival of heart failure patients by applying SMOTE as a method to address data imbalance and PSO as a feature selection technique within the Random Forest classification algorithm.

## METHODS

This research is conducted to examine the application of SMOTE and PSO in addressing dataset issues within the implementation of the Random Forest algorithm for the classification of heart failure patient survival. Below are the overall steps of the method used in this research, represented in the flowchart in Figure 1.

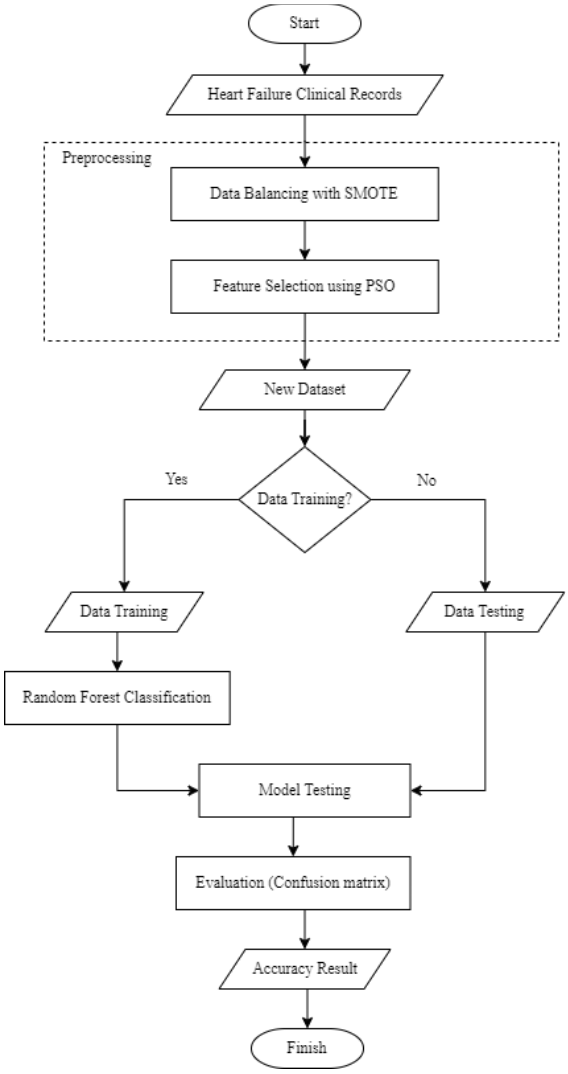


Figure 1. Flowchart of the Method Used

**Data Collection**

The data collection method employed in this research is a literature review. This literature review was conducted at the beginning of the study to deepen our understanding of the research problem and to review previous studies related to the topic under investigation. Based on the literature review, this study utilizes secondary data obtained from a dataset provider's website. The dataset used in this research is the Heart Failure Clinical Records Dataset, which was downloaded from the UCI Machine Learning Repository website. This dataset has been used in several previous studies to compare accuracy and algorithm performance.

**Data Analysis**

Data analysis will be conducted after obtaining the data. The initial step in data analysis is data preprocessing, which includes checking for missing values, data duplication, data balance, and feature selection. This step is performed to address data issues and prepare the data for the subsequent processes.

After conducting data checks and finding no missing values or data duplications, the preprocessing step continues with the data balancing process using SMOTE. The oversampling technique generates synthetic data for the minority class to balance the data between the two classes. The steps of the SMOTE process are as follows:

1. Identify the minority class in the dataset.
2. Select a sample from the minority class for oversampling.

3. Determine the number of synthetic data points to be generated.
4. Define the value of k-nearest neighbors. Based on the research conducted by [23], using k=5 to generate synthetic data.
5. Randomly select one example from the minority class.
6. Determine the KNN observations by sorting the distance of the selected example to all observations in the minority class using the Euclidean formula in Equation 1.

$$d_{(x,y)} = \sqrt{\sum_{i=i}^{n}(x_i - y_i)^2} = |x_i - y_i| \quad (1)$$

7. Generate synthetic data by determining the values of explanatory variables using Equation 2.

$$y_i = x_i + rand \times (w_i - x_i) \quad (2)$$

8. Repeat steps 5 to 7 until the desired amount of oversampling samples is achieved.

The next step involves performing feature selection using PSO by determining the gbest for each particle based on individual and swarm experiences. The steps of PSO process are as follows:
1. Initialize PSO parameters.
2. Calculate and evaluate the fitness values of each particle using Equation 3.

$$fitness = Accuracy = \frac{the\ number\ of\ correctly\ classified\ instance}{the\ number\ of\ instance} \quad (3)$$

3. Determine and update the values of pbest and gbest. Use Equation 4 for pbest and Equation 5 for gbest.

$$if\ (pos > pbest) : pbest = pos \quad (4)$$

$$if\ (pos > gbest) : gbest = pos \quad (5)$$

4. Update the velocity of particles using Equation 6 and update the position of particles using Equation 7 [24].

$$v_{id}^{t} = w \times v_{id}^{t-1} + (c_1 \times r_1 \times (p_{id}^{t} - x_{id}^{t})) + (c_2 \times r_2 \times (p_{gd}^{t} - x_{id}^{t})) \quad (6)$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t}\ ,\ d = 1, 2, \ldots, D. \quad (7)$$

5. Define the optimization criteria. Determine particle probabilities of 0 and 1 using the Sigmoid function (S) in Equation 8. Select features based on the calculations from Equation 9.

$$S = \frac{1}{1+e^{-v_{id}^{t}}} \quad (8)$$

$$x_{id}^{t} = \begin{cases} 1, if\ rand(0,1) < Sigmoid(S) \\ 0, otherwise \end{cases} \quad (9)$$

6. Repeat the process until the maximum iteration is reached. The selected attributes are obtained by taking the highest fitness value across all iterations.

After the preprocessing stage is completed, the next step is to divide the data into training and testing sets. The dataset is split into two portions with an 80% allocation for training data and 20% for testing data. Subsequently, data classification is carried out using the Random Forest algorithm through the ensemble of decision trees. This involves training on the available data samples and conducting a voting process for each class within the data samples.

The final step involves evaluating the classification results using a confusion matrix to determine the accuracy of the classification model. This step is essential for testing the model and calculating the accuracy based on the testing data. The accuracy value is calculated using the formula provided in Equation 10.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (10)$$

## RESULT AND DISCUSSION

### Data Collection Results

In this research, the data used is the Heart Failure Clinical Records Dataset, which was downloaded from the UCI Machine Learning Repository website. This dataset was initially developed by Davide Chicco and Giuseppe Jurman and later contributed to the data repository in 2020. The dataset comprises 299 rows of data with 12 attributes and 1 class attribute. The class attribute has two values, '0' representing the class with 'survived' status, and '1' for the class with 'dead' status. The number of data in class '0' is 203 (67.9%), and in class '1' there are 96 (32.1%) data. Below is the description of the dataset attributes in Table 1.

Table 1. Description of Attributes in the Heart Failure Clinical Records Dataset.

| No. | Attribute | Range | Attribute type |
|---|---|---|---|
| 1. | Age | [40,…,95] | Numeric |
| 2. | Anaemia | 0:No, 1:Yes | Biner |
| 3. | Creatinine phosphokinase | [23,…,7861] | Numeric |
| 4. | Diabetes | 0:No, 1:Yes | Biner |
| 5. | Ejection fraction | [14,…,80] | Numeric |
| 6. | High blood pressure | 0:No, 1:Yes | Biner |
| 7. | Platelets | [25.01,…,850.00] | Numeric |
| 8. | Serum creatinine | [0.50,…,9.40] | Numeric |
| 9. | Serum sodium | [114,…,148] | Numeric |
| 10. | Sex | 0:Female, 1:Male | Biner |
| 11. | Smoking | No:0, Yes:1 | Biner |
| 12. | Time | [4,…,285] | Numeric |
| 13. | Death event | 0:Survived, 1:Dead | Biner |

### Data Processing Results

### Data Balancing Results with SMOTE

Data balancing is conducted to address the issue of data imbalance within the dataset. The SMOTE method balances the data through oversampling, generating synthetic data based on KNN. The initial Heart Failure Clinical Records Dataset consisted of 299 data, with 203 data in class '0' (survived) and 96 data in class '1' (dead). The SMOTE method generates 107 synthetic data for class '1'. Consequently, the total data in the dataset after applying SMOTE is 406 data, with both class '0' and class '1' having 203 data each. Below is a sample of the dataset after undergoing the SMOTE process, as shown in Table 2.

Table 2. Dataset After SMOTE Process

| No | Age | Anaemia | CPK | …. | Smoking | Time | Death Event |
|---|---|---|---|---|---|---|---|
| 1 | 75 | 0 | 582 | …. | 0 | 4 | 1 |
| 2 | 55 | 0 | 7.861 | …. | 0 | 6 | 1 |
| 3 | 65 | 0 | 146 | …. | 1 | 7 | 1 |
| 4 | 50 | 1 | 111 | …. | 0 | 7 | 1 |
| 5 | 65 | 1 | 160 | …. | 0 | 8 | 1 |
| …. | …. | …. | …. | …. | …. | …. | …. |
| 402 | 57,87 | 0,36 | 413,18 | …. | 0,36 | 42,24 | 1 |
| 403 | 81,21 | 0,0 | 582,0 | …. | 0,0 | 132,87 | 1 |
| 404 | 53,66 | 0,63 | 1.539,49 | …. | 0,0 | 155,68 | 1 |
| 405 | 77,32 | 1,0 | 388,51 | …. | 0,0 | 24,39 | 1 |
| 406 | 51,97 | 0,09 | 794,95 | …. | 0,90 | 52,74 | 1 |

### Feature Selection Results with PSO

The implementation process of PSO for feature selection on the Heart Failure Clinical Records Dataset resulted in a reduction from 12 features to 10 relevant features being selected. Below are the selected features, represented with a value of 1, while the unselected features are represented with a value of 0, as shown in Table 3.

Table 3. Selected Features from the PSO Process

| No. | Attribute | Representation |
|---|---|---|
| 1. | Age | 1 |
| 2. | Anaemia | 1 |
| 3. | Creatinine phosphokinase | 1 |
| 4. | Diabetes | 0 |

| | | |
|---|---|---|
| 5. | Ejection fraction | 1 |
| 6. | High blood pressure | 1 |
| 7. | Platelets | 1 |
| 8. | Serum creatinine | 1 |
| 9. | Serum sodium | 0 |
| 10. | Sex | 1 |
| 11. | Smoking | 1 |
| 12. | Time | 1 |

**Data Splitting**

After completing the preprocessing stage, the next step is to divide the data into two parts, the training data and the testing data. Data splitting is performed using the splitting method provided by the sklearn library, with a data allocation proportion of 80% for training data and 20% for testing data. This division is carried out randomly, with a specified random state to ensure consistency in the randomization process.

**Data Mining Results**

The data mining stage is conducted by researching the Random Forest algorithm combined with the SMOTE and PSO methods. This stage consists of three data mining processes. Each process conducted in this research will be implemented using the Heart Failure Clinical Records Dataset. Subsequently, a comparison will be made based on the accuracy results obtained from each process that has been conducted. The first process involves conducting classification using the Random Forest algorithm on the Heart Failure Clinical Records Dataset without implementing the SMOTE method for class balancing and PSO for feature selection. The application of the Random Forest algorithm yields accurate results as shown in Table 4.

Table 4. Accuracy of the Random Forest Algorithm

| Algorithm | Accuracy |
|---|---|
| Random Forest | 88,33% |

The accuracy result of the Random Forest algorithm without implementing SMOTE and PSO on the Heart Failure Clinical Records Dataset is 88.33%. This result indicates that the classification using the Random Forest algorithm is quite good. However, there is room for improvement in the accuracy by implementing data balancing with SMOTE and feature selection with PSO.

The second process involves conducting classification using the Random Forest algorithm on the Heart Failure Clinical Records Dataset after applying the SMOTE method for data balancing. Below is the accuracy results achieved from the combination of SMOTE and Random Forest, as shown in Table 5.

Table 5. Accuracy of the Random Forest Algorithm on SMOTE-Processed Data

| Algorithm | Accuracy |
|---|---|
| Random Forest + SMOTE | 92,68% |

The accuracy achieved by applying the Random Forest algorithm with SMOTE results in a higher accuracy compared to the Random Forest algorithm without SMOTE, at 92.68%. This accuracy improvement represents a 4.35% increase over the previous accuracy result.

The third process involves conducting classification using the Random Forest algorithm while applying both SMOTE for data balancing and PSO for feature selection on the Heart Failure Clinical Records Dataset. You can view the accuracy results of this combination of SMOTE, PSO, and Random Forest in Table 6.

Table 6. Accuracy of the Random Forest Algorithm with SMOTE and PSO

| Algorithm | Accuracy |
|---|---|
| Random Forest + SMOTE + PSO | 93,9% |

The classification process on the Heart Failure Clinical Records Dataset using the Random Forest algorithm with SMOTE resulted in an accuracy of 92.68%. However, when combining SMOTE and PSO with the Random Forest algorithm, the accuracy increased to 93.9%. This represents an improvement of 1.22% in accuracy. Furthermore, when compared to the accuracy obtained from the Random Forest algorithm alone, the combination of SMOTE and PSO with the Random Forest algorithm resulted in a 5.57% difference in accuracy, indicating a significant improvement.

**Discussion**

This research implements the Random Forest classification algorithm with the data balancing method SMOTE and feature selection using PSO. The study aims to assess the performance of the Random Forest algorithm in classifying the Heart Failure Clinical Records Dataset when combined with SMOTE and PSO. The algorithm's performance in classifying the dataset can be determined by comparing the accuracy results of the Random Forest algorithm before and after combining it with SMOTE and PSO.

The implementation of SMOTE is carried out to address the issue of data imbalance in the Heart Failure Clinical Records Dataset by oversampling, which creates synthetic data based on the KNN values of the minority class. The minority class in this dataset is labeled '1' or 'dead' and it consists of 96 data. Meanwhile, the majority class labeled '0' or 'survived' has 203 data. To achieve class balance, data balancing between the classes is necessary, and this is achieved through oversampling using SMOTE. The application of SMOTE results in the generation of 107 synthetic data for the minority class. As a result, the SMOTE process produces a balanced dataset with a total of 406 data, evenly divided into 203 data for the minority class '1' and 203 data for the majority class '0'.

The feature selection method employed is the PSO method, which can select the best features from the dataset. The feature selection process is based on the highest fitness values generated by PSO, considering the best experiences of each particle and the swarm. The Heart Failure Clinical Records Dataset initially consisted of 12 attributes, and after feature selection with PSO, 10 selected attributes remained. These selected attributes include age, anemia, creatine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, sex, smoking, and time. The selected attributes are then used for the classification process to optimize the performance of the Random Forest algorithm.

This research records the accuracy results of each data mining process, including the Random Forest classification process, the Random Forest and SMOTE process, and the Random Forest with SMOTE and PSO process. The accuracy results are displayed in Table 7.

Table 7. Accuracy Results for Each Method

| Method | Accuracy |
|---|---|
| Random Forest | 88,33% |
| Random Forest + SMOTE | 92,68% |
| Random Forest + SMOTE + PSO | 93,9% |

Based on Table 7, it can be observed that there is an increase in accuracy for each method used. In the Random Forest classification process without using SMOTE and PSO, an accuracy of 88.33% is achieved. On the other hand, the application of SMOTE in the Random Forest classification process results in an accuracy of 92.68%. There is an improvement of 4.35% in accuracy due to the impact of data balancing. These accuracy results demonstrate that the combination of the Random Forest algorithm and SMOTE is capable of classifying the Heart Failure Clinical Records Dataset effectively and can enhance accuracy outcomes.

Meanwhile, in the proposed method, which involves the application of SMOTE and PSO in the Random Forest classification process, the highest accuracy of 93.9% is achieved. This result indicates a 5.57% improvement in accuracy compared to the application of the Random Forest algorithm without SMOTE and PSO. The increase in accuracy is attributed to the influence of data balancing and the attributes used in the classification process. SMOTE effectively balances the data with its oversampling technique, and PSO selects the best feature set based on the best experiences of its particles and the swarm. The accuracy comparison in this research is also conducted to demonstrate that the method applied in this study has advantages over previous research methods. Researchers compare the accuracy results obtained in this study with those of other studies based on the use of the same method or same dataset. The comparative results can be seen in Table 8.

Table 8. Comparison of Previous Research Accuracy

| Method | Dataset | Accuracy |
|---|---|---|
| Random Forest + SMOTE [16] | Heart Failure Clinical Records | 90% |
| Naive Bayes + PSO [25] | Heart Failure Clinical Records | 92,67% |
| Random Forest + Chi Square [26] | Heart Disease Statlog | 83,70% |
| Random Forest + SMOTE + PSO | Heart Failure Clinical Records | 93,9% |

Based on the accuracy comparison with previous research, it can be observed that the accuracy achieved in this study is superior to previous research methods. The key differentiator between this research and previous studies is the implementation of the SMOTE and PSO methods in the Random Forest classification process on the Heart Failure Clinical Records Dataset.

## CONCLUSION

The application of SMOTE to address the issue of data imbalance with the oversampling of the Heart Failure Clinical Records Dataset resulted in the addition of 107 data, balancing the data in both classes. The initial data consisted of 299 data, of which 96 were labeled '1' and 203 were labeled '0'. After SMOTE, the dataset contained a total of 406 data, achieving class balance. Subsequently, PSO was applied to dataset to select the most relevant features. The dataset contained 12 features before feature selection, and after the application of PSO, only 10 features remained. After the processes of SMOTE and PSO, the data was split into training and testing sets with an 80% and 20% ratio for implementation in the Random Forest classification algorithm. The selection of relevant features enhanced the performance of the Random Forest classification algorithm, making it more optimal.

The evaluation results using a confusion matrix, which includes the accuracy obtained from the Random Forest classification algorithm, yielded an accuracy of 88.33%. Subsequently, the application of the SMOTE and PSO methods to the Random Forest algorithm resulted in an accuracy of 93.9%. From this explanation, it can be concluded that the combination of applying SMOTE and PSO to the Random Forest classification algorithm successfully improved the accuracy by 5.57% in predicting the survival of heart failure patients. The implementation of SMOTE as a data balancing method and the selection of relevant features using PSO enhanced the performance of the Random Forest classification algorithm, leading to better accuracy.

## REFERENCES


[1] Suyanto, *Data Mining untuk Klasifikasi dan Klastering Data*. Bandung: Informatika Bandung, 2019.

[2] Han, J., Kamber, M., and Pei, J, *Data Mining: Data Mining Concepts and Techniques*. USA: Morgan Kaufmann, 2012, doi: 10.1109/ICMIRA.2013.45.

[3] Rohman, A., and Rochcham, M, "Model Algorithm C4.5 untuk Prediksi Penyakit Jantung," *Jurnal Neo Teknika*, vol. 4, no. 2, pp. 52–55, 2018, doi: 10.37760/neoteknika.v4i2.1228.

[4] Metra, M., and Teerlink, J. R, "Heart Failure," *The Lancet*, vol. 390, no. 10106, pp. 1981–1995, 2017.

[5] Purbianto, and Agustanti, D, "Analisis Faktor Risiko Gagal Jantung Di RSUD dr. H. Abdul Moeloek Provinsi Lampung," *Jurnal Keperawatan*, vol. XI, no. 2, pp. 194–203, 2015.

[6] Lippi, G., and Sanchis-Gomar, F, "Global Epidemiology and Future Trends of Heart Failure," *AME Medical Journal*, no. 5, vol. 15, pp. 1–6, 2020, doi: 10.21037/amj.2020.03.03.

[7] Rady, E. H. A., and Anwar, A. S, "Prediction of Kidney Disease Stages Using Data Mining Algorithms," *Informatics in Medicine Unlocked*, no. 15, pp. 1–7, 2019, doi: 10.1016/j.imu.2019.100178.

[8] Korzhakin, D. A., and Sugiharti, E, "Implementation of Genetic Algorithm and Adaptive Neuro Fuzzy Inference System in Predicting Survival of Patients with Heart Failure," *Scientific Journal of Informatics*, vol. 8, no. 2, pp. 251-257, 2021.

[9] Kurniawan, Y. I, "Perbandingan Algorithm Naive Bayes dan C.45 dalam Klasifikasi Data Mining," *Jurnal Teknologi Informasi Dan Ilmu Komputer*, no. 5, vol. 4, pp. 455–464, 2018, doi: 10.25126/jtiik.201854803.

[10] Singh, D., Choudhary, N., and Samota, J, "Analysis of Data Mining Classification with Decision tree Technique," *Global Journal of Computer Science and Technology*, vol. 13, no. 13, 2013.

[11] Kulkarni, V. Y., and Sinha, P. K, "Effective Learning and Classification Using Random Forest Algorithm," *International Journal of Engineering and Innovative Technolgy*, vol. 3, no. 11, pp. 267–273, 2014.

[12] Vijiyakumar, K., Lavanya, B., Nirmala, I., and Sofia Caroline, S, "Random Forest Algorithm for The Prediction of Diabetes," *International Conference on System, Computation, Automation and Networking, (ICSCAN)*, pp. 1–5, 2019, doi: 10.1109/ICSCAN.2019.8878802.

[13] Edric, and Tamba, S. P, "Prediksi Penyakit Gagal Jantung dengan Menggunakan Random Forest," *Jurnal Sistem Informasi Dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 176–181, 2022.

[14] Dittman, D. J., Khoshgoftaar, T. M., and Napolitano, A, "Is Data Sampling Required When Using Random Forest for Classification on Imbalanced Bioinformatics Data," *Advances in Intelligent Systems and Computing*, vol. 446, pp. 157–171, 2016, doi: 10.1007/978-3-319-31311-5_7.
[15] Sáez, J. A., Krawczyk, B., and Woźniak, M, "Analyzing The Oversampling of Different Classes and Types of Examples in Multi-class Imbalanced Datasets," *Pattern Recognition*, no. 57, pp. 164–178, 2016, doi: 10.1016/j.patcog.2016.03.012.
[16] Erlin, Desnelita, Y., Nasution, N., Suryati, L., and Zoromi, F, "Dampak SMOTE terhadap Kinerja Random Forest Classifier Berdasarkan Data No Seimbang," *Matrik: Jurnal Manajemen, Teknik Informatika, Dan Rekayasa Komputer*, vol. 21, no.3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.
[17] Zhu, T., Lin, Y., amd Liu, Y, "Synthetic Minority Oversampling Technique for Multiclass Imbalance Problems," *Pattern Recognition*, no. 72, pp. 327–340, 2017, doi: 10.1016/j.patcog.2017.07.024.
[18] Xue, B., Zhang, M., Member, S., and Browne, W. N, "Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach," *IEEE Transactions on Cybernetics*, pp. 1–16, 2012.
[19] Aghdam, M. H., and Heidari, S, "Feature Selection using Particle Swarm Optimization in Text Categorization," *JAISCR*, no. 5, vol. 4, pp. 231–238, 2015, doi: 10.1007/978-81-322-1985-9_2.
[20] Lin, S. W., Ying, K. C., Chen, S. C., and Lee, Z. J, "Particle Swarm Optimization for Parameter Determination and Feature Selection of Support Vector Machines," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008, doi: 10.1016/j.eswa.2007.08.088.
[21] Ramanda, K., and Carolina, I, "Seleksi Fitur Algorithm Neural Network Menggunakan Particle Swarm Optimization Untuk Memprediksi Kelahiran Prematur," *Kilat*, vol. 6, no. 2, pp. 106–111, 2017, doi: 10.33322/kilat.v6i2.134.
[22] Lubis, M. R, "Method Hybrid Particle Swarm Optimization - Neural Network Backpropagation untuk Prediksi Hasil Pertandingan Sepak Bola," *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, vol. 1, no. 1, pp. 71, 2017, doi: 10.30645/j-sakti.v1i1.30.
[23] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W, "SMOTE : Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[24] Kennedy, J., and Eberhart, R, "Particle Swarm Optimization," *Proceedings of ICNN'95*, vol. 4, pp. 1942–1948, 1995.
[25] Novaldy, F., and Herliana, A, "Penerapan PSO pada Naive Bayes untuk Prediksi Harapan Hidup Pasien Gagal Jantung," *Jurnal Responsif: Riset Sains Dan Informatika*, vol. 3, no. 1, pp. 37–43, 2021, https://doi.org/10.51977/jti.v3i1.396.
[26] Jabbar, M. A., Deekshatulu, B. L., and Chandra, P, "Prediction of Heart Disease Using Random Forest and Feature Subset Selection," *Advances in Intelligent Systems and Computing*, vol. 424, pp. 187–196, 2016, https://doi.org/10.1007/ 978-3-319-28031-8.