Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 1 Iss. 1 July 2023, pp: 34-40
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Implementation of Support Vector Machine and Random Forest for Heart Failure Disease Classification

Astriana Rahmah[1*], Nurhafiza Sepriyanti[2], Muhammad Hafis Zikri[3], Isnani Ambarani[4], Muhammad Yusuf bin Shahar[5]

[1,2,3,4]Department of Information Systems Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[5]Malaysia Islamic Science University, Malaysia

E-Mail: [1]12050320398@students.uin-suska.ac.id,
[2]12050321863@students.uin-suska.ac.id, [3]1205030313115@students.uin-suska.ac.id,
[4]12050327049@students.uin-suska.ac.id, [5]yusufshaharmat@gmail.com

Received March 19th 2023; Revised Jun 6th 2023; Accepted Jul 10th 2023
Corresponding Author: Astriana Rahmah

Abstract

Heart failure is a life-threatening disease and its management should be considered a global public health priority. The use of data mining in data processing operations to identify existing patterns and identify the information stored in them. In this study, researchers classify using two algorithms for comparison of algorithms, namely Random Forest (RF) and Support Vector Machine (SVM). The purpose of this study is to find patterns in finding the best accuracy for the 2 algorithms. The results of this study obtained an accuracy of 81.51%. with a Hold Out of 60 : 40% on the SVM algorithm, while an accuracy of 83.33 % with a Hold Out of 9 0 : 1 0% on the R F algorithm . From these results it can be seen that the highest accuracy value is obtained at RF making the best algorithm compared to the SVM algorithm.

Keywords**:** Classification, Heart failure, Random Forest, Support Vector Machine

1. INTRODUCTION

The heart is the first organ to form in the body child and is located in the mother's womb . The fetal heart begins to beat for the first time in the sixth week of pregnancy, and other organs are built around the heart and blood vessels in the fetus's body. The heart beats harder and faster in situations that require a lot of effort, such as running, climbing or sports such as soccer. In other words, the heart is the source of life and vitality for humans and living beings, and perhaps because of that it is believed that the soul is kept in the heart [1].

Heart failure is a life-threatening disease and its management should be considered a global public health priority. Currently around 26 million people in the world suffer from this disease. The outlook for such patients is poor, with survival rates worse than those of colon, breast or prostate cancer [2]. Currently, the number of heart failure patients continues to increase due to population growth and aging [3].

By using data mining techniques can detect disease in the first stage. Mining data technique is the method of approach. It is used to examine large amounts of information [4]. Previous research was conducted by Farzana Taznim and SU Habiba (2021) [5] using RF and PCA models which produced the best accuracy of 92.85% compared to other algorithms. This model can predict the probability of coronary heart disease. Then, a study by Ritaban Mitra and T. Rajendran (2022) resulted in an average accuracy of 94.61% Random Forest (RF) and 93.91% in the Support Vector Machine (SVM) algorithm. The Random Forest algorithm is able to predict stroke efficiently, because it has the best accuracy [6].

Support Vector Machine (SVM) has been instrumental in many new issues in various fields. Recently, SVM has become the most popular tool for data classification and data mining problems [7]. SVM is one of the most powerful machine learning algorithms that has been widely used in pattern recognition since the early 1990's [8]. SVM is a set of supervised learning methods that can applied to classification and regression problems [9]. RF is a powerful machine learning classifier that relatively unknown in remote sensing of the earth and has not been thoroughly evaluated against traditional pattern recognition techniques by the remote sensing community [10]. SVM offers many unique advantages for solving small, nonlinear, and high-dimensional pattern recognition problems [11]. While RF has several advantages compared to other classification methods. Can use continuous and categorical data sets, easy to parameterize, sensitive to overfitting , handles outliers in training data well, and calculates supporting information such as misclassification and variables [12].

In this study the authors classify using two algorithms for comparison of algorithms namely RF and SVM. By using these two algorithms, researchers will make comparisons between these methods to determine the best algorithm to be used on the data. The data set used is taken from the Kaggle page published in 2021. Classification methods that have been carried out and reviewed in various previous literature, namely the RF and SVM classification methods, give good and satisfactory results. Based on this, this study applies the RF and SVM algorithms to look for patterns to find the best accuracy of the two algorithms in classifying heart failure data.

2. MATERIALS AND METHODS

The flow of this research can be seen in Figure 1. At the data collection stage, the dataset used was the Heart Failure Clinical Records data taken from the Kaggle page which was published in 2021. Then the data preprocessing stage was carried out using data normalization . with minmax technique . After that the dataset will be divided according to the training data and test data using the Hold Out technique of 90:10%, 80:20%, 70:30% and 60:40%. Furthermore, the classification is carried out with the SVM and RF algorithms. Then the analysis results obtained can be seen in Figure 1.

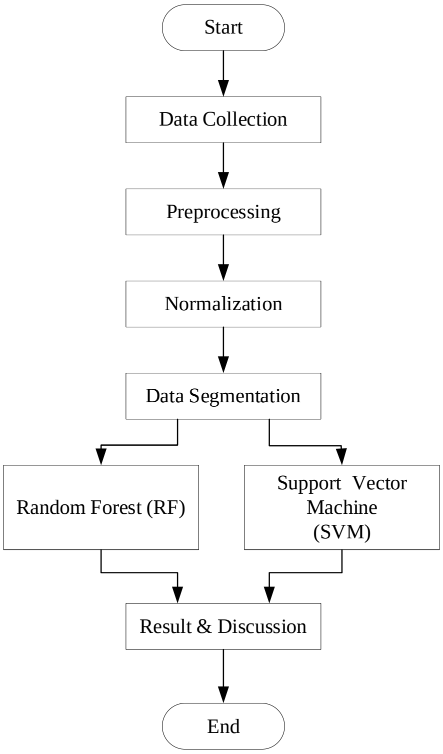

Figure 1. Research Methodology

2.1 Fail Heart

A condition when the heart is unable to pump blood based on the needs of the tissues is called heart failure [13]. In addition, heart failure is also one of the clinical health problems associated with death, morbidity, mortality, and health costs. In some European countries almost 2% direct cost of heart failure from the total health budget [14].

2.2 Normalization

Data normalization is one of the techniques at the preprocessing stage so that data can be mapped in the range 0 and 1. There are several data normalization methods, namely, min-max, Z-score, and feature scaling. In this study, the normalization method used is min-max normalization. Min-max normalization is used to solve the problem of values between features that are too far apart. So the formula for normalizing the data is as follows [15].

$$\mathrm{N}' = \frac{\mathrm{N}-\min(\mathrm{n})}{\max(\mathrm{n})-\min\,(\mathrm{n})} \tag{1}$$

2.3 Hold Out

In share training data (training) and test data (testing) can done with validation holdout that is , the data is divided into two parts with proportions that have determined by the researcher . After valuable data is obtained namely training data and test data, researchers will train a classification model with training data and testing it with test data. The data sharing technique known as hold out ie division of the dataset into two data based on comparison certain. On research This determine value of 70% for training data and 30% for test data [16].

2.4 Classification

Classification is a process of grouping the same objects/units and separating different objects/units from one another. The process of grouping data, namely the output parameters of feature extraction, based on the similarity of data features. Classification is also a process that groups data together with other data that has similarities in a characteristic [17].

The classification algorithm uses the training data to create a model. The created model is then used to predict new unknown data class tags. However, the principle of each algorithm is the same, namely conducting training in such a way that in the end the model is trained. can accurately predict the output class label of each input vector [18].

2.5. Support Vector Machine (SVM)

A method that analyzes data and identifies patterns in classification is the definition of the SVM algorithm. SVM has the advantage of being able to identify various hyperplanes that determine the boundaries between two different classes. However, SVM has a weakness in choosing the appropriate function. The selection of functions and parameter settings has an impact on the SVM on the results of classification accuracy [19] . SVM belongs to supervised learning , namely analyzing data and finding patterns for the classification process. Before being used for classification, text is required to be converted into vectors.

Equality Support Vector Machine (SVM):

$$\mathrm{f}\,(\mathrm{x}) = \mathrm{W}^{\mathrm{t}}\, \phi(\mathrm{x}) + \mathrm{b} \tag{1}$$

Information:

b : Biased
$\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots., \mathrm{x}_\mathrm{D})$: Input variable
$\mathrm{W} = (\mathrm{w}_0, \mathrm{w}_1, \ldots., \mathrm{w}_\mathrm{D})$: Weight parameter
$\phi(\mathrm{x})$: Feature transformation function

2.6. Random Forest (RF)

RF is approach learning ensemble developed by Breiman . _ RF is method learning machine new purpose _ finish problem classification and regression [20]. There are three necessary parameters emphasized in RF modeling . First is n_estimator . This means total trees in forest . Second is max_features which refers to the count maximum features in the tree . Third is min_samples_leaf This is minimum sample size for leaf nodes . Max_features related strength prediction every tree and power correlation between both . In increasing this parameter also expands ability prediction of each tree and strength correlation [21]. Classification process using RF ie with count attribute entropy as determinant level impurity and value Information Gai n. With use formula equation 1 can do calculation entropy value , meanwhile For count mark Gain Information use equation 2 [22].

Following formula For Random forests (RF):

$$\text{Entropy (Y)} = -\sum_{\mathrm{i}} p\,(\,c \mid Y\,)\; log^{\,2} p\,(\,c \mid Y\,) \tag{2}$$

Description:

Y : set case
P(c|Y) : Proportion mark Y to class c.

$$\text{Information gains (Yes) = Entropy (Y) = } -\sum_{\mathrm{v}\varepsilon \mathrm{Values(a)}} \frac{|\mathrm{Y_v}|}{|\mathrm{Y_a}|} \text{Entropy } (\mathrm{Y_v}) \tag{3}$$

Information:

Values(a) : Set values case a.
$\mathrm{Y_v}$: related subclass Y class v to class a.
$\mathrm{Y_a}$: Whole appropriate value _ to class a.

The big advantage of Random Forest is that it can be used for classification and regression problems, which are most of the learning systems . Because classification is sometimes considered a building block of machine learning [23] . Advantages of Random Forest has better results compared to other individual models because algorithm Random Forest uses decisions a tree that has no correlation [24] .

3. RESULTS AND ANALYSIS

3.1 Data Collection

Research is business Which carried out systematically with all stages to get a scientific answer from an answer [2 5] . The dataset used is a heart failure clinical record dataset taken from the Kaggle page which will be published in 2021. The data consists of 299 data with 13 attributes on heart failure patient data. The dataset used in this study is in Table 1.

Table 1. Preliminary Data

Age	Anemia	Creatinine Phosphokinase	Diabetes	Ejection Fraction	High Blood Pressure	Platelets	…..	Death Events
75	0	582	0	20	1	265000	…..	1
55	0	7861	0	38	0	263358	…..	1
65	0	146	0	20	0	162000	…..	1
50	1	111	0	20	0	210000	…..	1
65	1	160	1	20	0	327000	…..	1
90	1	47	0	40	1	204000	…..	1
75	1	246	0	15	0	127000	…..	1
60	1	315	1	60	0	454000	…..	1
65	0	157	0	65	0	263358	…..	1
80	1	123	0	35	1	388000	…..	1
…..	…..	…..	…..	…..	…..	…..	…..	…..
50	0	196	0	45	0	395000	…..	0

3.2 Data Normalization

After the data preprocessing process is carried out , then the data is normalized use normalization Min-Max , which is in the range 0-1 [26] . Normalization of data carried out in research This can seen in Table 2 as following .

Table 2. Data Normalization

Age	Anemia	Creatinine Phosphokinase	Diabetes	Ejection Fraction	High Blood Pressure	Platelets	…..	Death Events
0.636	0	0.0713	0	0.090	1	0.290	…..	1
0.272	0	1	0	0.363	0	0.288	…..	1
0.454	0	0.015	0	0.090	0	0.165	…..	1
0.181	1	0.011	0	0.090	0	0.224	…..	1
0.454	1	0.017	1	0.090	0	0.365	…..	1
0.909	1	0.003	0	0.393	1	0.216	…..	1
0.636	1	0.028	0	0.015	0	0.123	…..	1
0.363	1	0.037	1	0.696	0	0.519	…..	1
0.454	0	0.017	0	0.772	0	0.288	…..	1
0.727	1	0.012	0	0.318	1	0.439	…..	1
…..	…..	…..	…..	…..	…..	…..	…..	…..
0.469	0	0.448	0.123	0.657	1	1	…..	0

Can observed in Table 2 **.** explain that results dataset classification consists of two classes , namely 0 indicates No fail heart and 1 shows fail heart .

3.3 Data Sharing

In this stage to obtain the results of the accuracy level, a method is needed, namely Hold Out . In dividing the data using the Hold Out method, a comparison of training and testing was obtained , namely 90%, 80%, 70% and 60%.

3.4 Algorithm Support Vector Machine (SVM)

The research classification process uses two algorithms, one of the algorithms used is SVM using Hold Out distribution data of 90:10%, 80:20%, 70:30% and 60:40%. The following results of the evaluation comparison on the SVM algorithm can be seen in Figure 2.

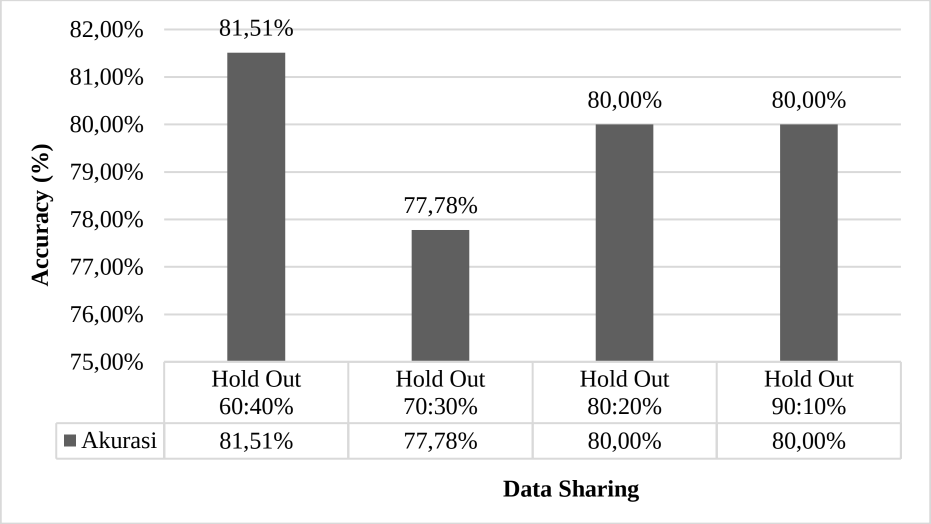

Figure 2. Comparison Evaluation SVM Algorithm

From Figure 2, stated that data distribution with use Hold Outs 60% more effective on the SVM algorithm compared with Hold Out other with level accuracy of 81.51%.

3.5 Algorithm Random Forest (RF)

Besides SVM algorithm , research it also uses RF algorithm . On the algorithm This done with use same data distribution ie Hold Out 90:10% to 60:40%. Following is results comparison every Hold Out on the RF algorithm can seen in Figure 3.

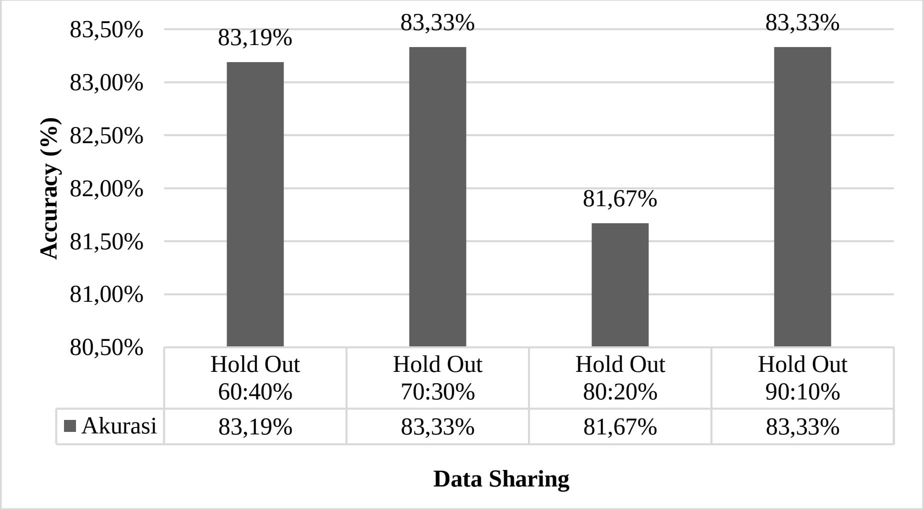

Figure 3. Comparison Evaluation Algorithm Random Forest

Based on results diagram Figure 3, states that data distribution with use Hold Outs 90:10% more effective on the RF algorithm compared with Hold Out other.

3.6 Comparison Second Algorithm

At stage This For compare second algorithm SVM and RF researchers use RapidMiner Studio software. Then can obtained that from comparison for the two algorithms This accuracy best is in the algorithm Random Forest with accuracy best 83.33% at 90:10% Hold Out.

Table 3. Accuracy , Precision and Recall Algorithm SVM

SVM	Accuracy	Precision	Recall
Hold Out 60:40%	81.51%	79.80%	97.53%
Hold Out 70:30%	77.78%	76.62%	96.72%
Holdouts 80:20%	80.00%	79.59%	95.12%
Holdouts 90:10%	80.00%	79.17%	95.00%

Table 4. Accuracy , Precision and Recall Algorithms Random Forest

Random Forest	Accuracy	Precision	Recall
Hold Out 60:40%	83.19%	82.11%	96.30%
Hold Out 70:30%	83.33%	81.94%	96.72%
Holdouts 80:20%	81.67%	80.00%	97.56%
Holdouts 90:10%	83.33%	80.00%	100.00%

Based on tables 3 and 4 known in the SVM algorithm produce accuracy Lowest namely 77.78% , then Precision which is 76.62%, and Recall namely 96.72% at Hold Out 70:30 and accuracy highest namely 81.51%, Precision namely 79.80, and Recall namely 97.53% at Hold Out 60:40%, meanwhile RF algorithm generates accuracy Lowest i.e. 81.67 , Precision namely 81.67%, and Recall which is 97.56% at Hold Out 80:20 % and accuracy the highest is 83.33% , Precision is 82.11%, and Recall is 96.30% at Hold Out 90:10%.

4. CONCLUSIONS

From the results of this study using a dataset of heart failure clinical records with two SVM algorithms with RF obtained an accuracy of 81.51%. with 60:40% Hold Out on the SVM algorithm, while with 83.33% accuracy Hold Out 90:10% on RF algorithm . From these results it can be seen that the highest accuracy value is obtained on RF making the algorithm the best compared to the SVM algorithm. Thus it can be concluded that the RF algorithm is quite good at classifying heart failure.

REFERENCES

[1] AK Faieq and MM Mijwil, "Prediction of heart diseases using support vector machines and artificial neural networks," Indones. J.Electr. Eng. Comput. Sc., vol. 26, no. 1, pp. 374–380, 2022.

[2] P. Ponikowski et al., “Heart failure: preventing disease and death worldwide,” ESC Hear. Files, vol. 1, no. 1, pp. 4–25, 2014.

[3] A. Groenewegen, FH Rutten, A. Mosterd, and AW Hoes, “Epidemiology of heart failure,” Eur. J. Heart Fail., vol. 22, no. 8, pp. 1342–1356, 2020.

[4] SP Shaji, "Prediction and diagnosis of heart disease patients using data mining technique," in 2019 international conference on communication and signal processing (ICCSP), 2019, pp. 848–852.

[5] A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection

[6] Efficient Prediction of Stroke Patients Using Random Forest Algorithm in Comparison to Support Vector Machine

[7] VF Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and JP Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” ISPRS J. Photogramm . Remote Sens., vol. 67, pp. 93–104, 2012.

[8] OL Mangasarian, “Data mining via support vector machines,” in IFIP Conference on system modeling and optimization, 2001, pp. 91–112.

[9] S. Karamizadeh, SM Abdullah, M. Halimi, J. Shayan, and M. javad Rajabi, “Advantage and drawbacks of support vector machine functionality,” in 2014 international conference on computer, communications, and control technology (I4CT) , 2014, pp. 63–65.

[10] A. Shmilovici, “Support vector machines,” in Data mining and knowledge discovery handbook, Springer, 2009, pp. 231–247.

[11] J. Du, Y. Liu, Y. Yu, and W. Yan, “A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms,” Algorithms, vol. 10, no. 2, p. 57, 2017.

[12] N. Horning, “Random Forests: An algorithm for image classification and generation of continuous field data sets,” in Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan, 2010, vol . 911, pp. 1–6.

[13] SA Dick, R. Zaman, and S. Epelman, “Using high-dimensional approaches to probe monocytes and macrophages in cardiovascular disease,” Front. Immunol., vol. 10, p. 2146, 2019.

[14] AP Lumi, VFF Joseph, and NCI Polii, "Cardiac Rehabilitation in Patients with Chronic Heart Failure," J. Biomedicine JBM, vol. 13, no. 3, pp. 309–316, 2021.

[15] Performance Comparison of the K-Means Method for Classification in Diabetes Patients Using Two Normalization Methods

[16] P. Purwono, A. Wirasto, and K. Nisa, "Comparison of Machine Learning Algorithms for Classification of Drug Groups," SISFOTENIKA, vol. 11, no. 2, pp. 196–207, 2021.

[17] M. Mayasari, D. Iskandar Mulyana, M. Betty Yel, and S. Higher Computer Science Cipta Karya Informatika Jl Raden, "Comparison of Classification of Rhizome Plant Types Using Principal Component Analysis, Support Vector Machine, K-Nearest Neighbor and Decision Tree," J.Tek. inform. Kaputama, vol. 6, no. 2, 2022.

[18] Garcia-Carretero, Rafael, et all. 2020. Use of a K-Nearest Neighbors Model to Predict The Development of Type 2 Diabetes Within 2 Years in An Obese, Hypertensive Population. International Federation for Medical and Biological Engineering.

[19] D. A. Kristiyanti, "Analisis Sentimen Review Produk Kosmetik menggunakan Algoritma Support Vector Machine dan Particle Swarm Optimization sebagai Metode Seleksi Fitur," SNIT 2015, vol. 1, no. 1, pp. 134–141, 2015.

[20] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[21] J. Peters, NEC Verhoest, R. Samson, P. Boeckx, and B. De Baets, "Wetland vegetation distribution modeling for the identification of constraining environmental variables," Landscape Ecol., vol. 23, no. 9, pp. 1049–1065, Sp. 2008, doi: 10.1007/s10980-008-9261-4.

[22] K. Schouten, F. Frasincar, and R. Dekker, "An information gain-driven feature study for aspect-based sentiment analysis," in International Conference on Applications of Natural Language to Information Systems, 2016, pp. 48–59.

[23] B. An and Y. Suh, "Identifying financial statement fraud with decision rules obtained from Modified Random Forest," Data Technol. Appl., vol. 54, no. 2, pp. 235–255, 2020.

[24] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," IEEE access, vol. 7, pp. 180235–180243, 2019.

[25] F. Akbar, HW Saputra, AK Maulaya, MF Hidayat, and R. Rahmaddeni, "Implementation of the C4 Decision Tree Algorithm. 5 and Support Vector Regression for Prediction of Stroke: Implementation of Decision Tree Algorithm C4. 5 and Support Vector Regression for Stroke Disease Prediction," MALCOM Indonesia. J. Mach. learn. Comput. Sc., vol. 2, no. 2, pp. 61–67, 2022.

[26] N. Sepriyanti, R. S. Nahampun, M. H. Zikri, I. Ambarani, and A. Rahmadeyan, "Penerapan K-Means Clustering Untuk Mengelompokkan Tingkat Kemiskinan di Provinsi Riau: Implementation of K-Means Clustering to Group Poverty Levels in Riau Province," in SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat, 2022, vol. 1, no. 1, pp. 59–65.