PREDICTION OF SURVIVAL OF HEART FAILURE PATIENTS USING RANDOM FOREST

Sri Rahayu[1*]; Jajang Jaya Purnama[2]; Achmad Baroqah Pohan[3]; Fitra Septia Nugraha[4]; Siti Nurdiani[5]; Sri Hadianti[6]

[1, 4, 5] Informatics Engineering Study Program; [2, 6] Information Systems Study Program
STMIK Nusa Mandiri
www.nusamandiri.ac.id
srirahayu.rry@nusamandiri.ac.id; jajang.jjp@nusamandiri.ac.id; fitra.fig@nusamandiri.ac.id;
siti.sxd@nusamandiri.ac.id; sri.shv@nusamandiri.ac.id

[3] Computer Technology Study Program
Universitas Bina Sarana Informatika
www.ubsi.ac.id
achmad.abq@bsi.ac.id
(*) Corresponding Author

Abstract— Human survival, one of the roles that is controlled by the heart, makes the heart need to be guarded and be aware of its damage. Cardiovascular breakdown is the last phase of all coronary illness. The problem of the number of deaths caused by heart failure requires a survival predictor tool. The patient electronic clinical record apparatus is accessible to gauge manifestations, body highlights and clinical research center test esteems that can be utilized to perform biostatistical dissects pointed toward featuring examples and connections that are not recognized by clinical specialists. AI is an answer for have the option to foresee tolerant endurance from the information created and to have the option to recognize the most significant components among those remembered for their clinical records. With data mining techniques used in the available history data, namely the Heart Failure Clinical Records dataset of 299 instances on 13 features using the Random Forest algorithm, Decision Tree, KNN, Support Vector Machine (SVM), Artificial Neural Network and Naïve Bayes with resample and Synthetic Minority Oversampling Technique (SMOTE) sampling techniques. The highest accuracy with the resample sampling technique in the random forest is 94.31% and the SMOTE technique used in the random forest produces an accuracy of 85.82% higher than other algorithms. The example framed can anticipate the endurance of cardiovascular breakdown patients.

Keywords: data mining, heart failure, random forest, resampling, SMOTE

Abstrak—*Keberlangsungan hidup manusia yang salah satu peran pentingnya dikendalikan oleh jantung, membuat jantung perlu dijaga dan diwaspadai kerusakannya. Permasalahan mengenai banyaknya yang meninggal yang diakibatkan oleh gagal jantung perlu adanya alat prediksi keberlangsungan hidupnya. Alat rekam klinis elektronik pasien dapat diakses untuk mengukur manifestasi, sorotan tubuh, dan nilai uji pusat penelitian klinis yang dapat digunakan untuk melakukan pembedahan biostatistik yang menunjukkan contoh dan hubungan yang tidak dikenali oleh spesialis klinis. AI adalah jawaban karena memiliki pilihan untuk meramalkan ketahanan toleran dari informasi yang dibuat dan memiliki pilihan untuk mengenali komponen paling signifikan di antara yang diingat untuk catatan klinis mereka. Dengan teknik data mining yang digunakan pada data history yang tersedia yaitu dataset Heart Failure Clinical Records sebanyak 299 instance pada 13 feature menggunakan algoritma Random Forest, Decision Tree, KNN, Support Vector Machine, Artificial Neural Network dan Naïve Bayes dengan teknik sampling resample dan Synthetic Minority Oversampling Technique (SMOTE) menghasilkan akurasi paling tinggi dengan teknik sampling resample pada random forest yaitu* 94.31 *% dan teknik SMOTE yang digunakan pada random forest menghasilkan akurasi 85.82% lebih tinggi dari algoritma lain. pola yang terbentuk bisa memprediksi keberlangsungan hidup pasien gagal jantung.*

***Kata Kunci**: data mining, gagal jantung, random forest, resampling, SMOTE*

INTRODUCTION

The continuity of human life can never be separated from the organs that work continuously.

The heart is the main organ of the human body because of its very important task, namely pumping blood and distributing it throughout the body and later it is blood that carries oxygen and nutrients needed by the human body. The heart is a vital organ and is the last line of defense for life other than the brain. This pulse in the heart cannot be controlled by humans. Heart rate usually refers to the amount of time the heartbeat takes per unit of time (Rozie, Hadary, & Wigyarianto, 2016).

Heart disease is the number one killer in the world. Each year, more than 2 million Americans die from heart disease / stroke (Frieden & Berwick, 2011). Heart disease is a disorder that occurs in the large blood vessel system, causing the heart and blood circulation to not function properly. Diseases related to the heart and blood vessels include: heart failure, coronary heart disease, and rheumatic heart disease (Aeni, Santosa, & Supriyanto, 2014). Heart failure is the final stage of all heart disease and is the cause of increased morbidity and mortality in cardiac patients (Imaligy, 2014). The incidence of heart failure will increase in the future due to increasing life expectancy and the development of myocardial infarction treatment therapy resulting in improved life expectancy of patients with decreased heart function (Hamzah, 2016).

Problems that arise include the danger and death of heart failure when it is suffered by the human body (Frieden & Berwick, 2011), then this needs to be predicted so that it can be known in advance and treated or therapy as soon as possible to reduce mortality or prolong patient survival. The medical record tool can quantify side effects, body highlights, and clinical research facility test esteems, which can be utilized to perform biostatistical analyzes but to highlight patterns and correlations not detected by medical doctors (Chicco & Jurman, 2020). With the help of technology, the data from the biostatistical analysis can be processed with data mining techniques to form patterns of correlation between data from existing historical data so that it can make a prediction tool if implemented from the pattern formed earlier. Machine Learning specifically, can anticipate tolerant endurance from its information and can recognize the most significant components remembered for their clinical records (Chicco & Jurman, 2020).

This research is not the first to be conducted, other researchers have also done the same thing, here is the main reference paper with the following results :

Table 1. Main Reference Paper

Research Title	Method & Result Accuracy
Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone	Random forests : 74% Decision tree : 73,7% Gradient boosting : 73,8% Linear regression : 73% One rule : 72,9% Artificial neural network : 68% Naïve bayes : 69,6% SVM radial : 69% SVM linear : 68,4% k-nearest neighbors :62,4%

Source : (Chicco & Jurman, 2020)

In Table 1, we can see that in the paper entitled "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone" Researchers used 10 machine learning methods to predict the survival of heart failure patients with the attributes used, namely: from serum creatinine and ejection fraction, the highest accuracy is by using Random Forest which is 74% (Chicco & Jurman, 2020).

In another study, Tanvir Ahmad, et al used Cox regression to model mortality contemplations, age, launch division, serum creatinine, serum sodium, sickliness, platelets, creatinine phosphokinase, circulatory strain, sex, diabetes and smoking status as expected supporters of mortality. The Kaplan Meier plot was utilized to contemplate general examples of endurance demonstrating a high mortality force in the good 'ol days and afterward expanding step by step until the finish of the examination. The martingale buildup was utilized to evaluate the utilitarian type of the variable. The outcomes approved the incline of the computational adjustment and the capacity to separate between models through bootstrapping. For a graphical forecast of the likelihood of endurance, a nomogram is created. Age, renal brokenness, pulse, launch portion and paleness were discovered to be noteworthy danger factors for death among cardiovascular breakdown patients (Ahmad, Munir, Bhatti, Aftab, & Raza, 2017).

In this study we conducted a classification to predict the survival of heart failure patients (Chicco & Jurman, 2020) with data mining techniques using several classification algorithms in order to compare which algorithm is more suitable for use in the dataset used. If in the previous study the attributes of serum_creatinine and serum_sodium were used, in this study we used all the attributes, namely age, anemia, creatinine, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time with class death_event and preprocessing. data with resampling and SMOTE because the data used is imbalance data, this technique is a novelty because

it produces better accuracy than before on the same dataset.

MATERIALS AND METHOD

To conduct a research, of course, the dataset is the main material that will be processed in such a way using an algorithm. The dataset used in this research is Heart Failure Clinical Records data taken from the UCI repository website. The dataset was published in 2020 with 13 features, namely: age, anemia, creatinine, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time, with 299 instances of class death_event with the following information :

Table 2. Description of Dataset

No	Features	Description
1	*age*	patient age (years)
2	*anaemia*	decrease in red blood cells or hemoglobin
3	*creatinine*	CPK enzyme levels in the blood (mcg/L)
4	*diabetes*	if the patient has diabetes
5	*ejection_fraction*	the percentage of blood that leaves the heart with each contraction
6	*high_blood_pressure*	if people with hypertension
7	*platelets*	platelets in the blood (kiloplatelet / mL)
8	*serum_creatinine*	serum creatinine level in the blood (mg /dL)
9	*serum_sodium*	serum sodium level in the blood (mEq/L)
10	*sex*	Gender: female or male
11	*smoking*	if the patient smokes or not
12	*time*	follow-up period (days)
13	*death_event*	if the patient dies during follow-up

Source : (Chicco, 2020)

Ekperimen yang dilakukan pada dataset ini diantaranya menggunakan metode klasifikasi Random Forest (Wager & Athey, 2018) which is one of the methods used for classification and regression. This method is an ensemble of learning methods using a decision tree as a base classifier that is built and combined (Primajaya & Sari, 2018).

As a comparison, the Decision Tree algorithm is also used, which is a decision tree where each branch shows a choice among a number of alternative choices, and each leaf shows the selected decision (Setiawati, Taufik, Jumadi, & Z, 2016).

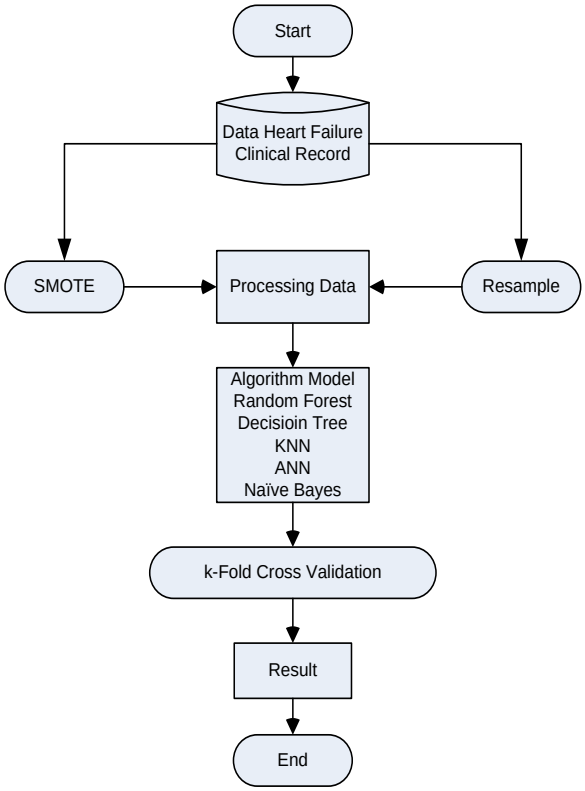
In addition, the Nearest Neighbor algorithm is also used, which is a classification algorithm based on analogy, which compares the test data with training data that is close to and has similarities to the test data (Sartika & Sensuse, 2017), also SVM classification method that works by searching for the hyperplane with the biggest hyperplane edge, which is the limit line isolating the information between classes. Edge is the separation between the hyperplane and the nearest information in each class. The information nearest to the hyperplane in each class is known as the support vector (Somantri, Wiyono, & Dairoh, 2016).

In addition to the four algorithms above, the Artificial Neural Network algorithm is also applied which is a non-linear statistical data modeling tool and can be used to model complex relationships between input and output to find data patterns (Syukri & Samsuddin, 2018) and Naïve Bayes is a basic likelihood order that figures a lot of probabilities by including the frequencies and worth blends from a given dataset. The calculation utilizes the Bayes hypothesis and accepts all credits are autonomous or not related given by values on class factors (Manalu, Sianturi, & Manalu, 2017) & (Aninditya, Hasibuan, & Sutoyo, 2019).

However, before classification is carried out, SMOTE preprocessing is carried out first, the use of the SMOTE (Synthetic Minority Over-Sampling Technique) technique produces good and effective results for handling classimbalance that is over-fitting in the over-sampling technique process for minority classes (positive)(Putri, 2017). SMOTE (Synthetic Minority Oversampling) is an oversampling approach that synthetically generates instances by randomly selecting instances from the minority class and using the interpolation method to generate instances between the selected point and adjacent instances (Shelke, Deshmukh, & Shandilya, 2017).

In addition to the SMOTE technique in preprocessing, the Resample technique was also carried out in this study, this method works by diminishing the larger part class test. This decrease should be possible arbitrarily for this situation called arbitrary undersampling or it very well may be finished utilizing some measurable information for this situation called data undersampling. Some are educated that the undersampling strategy and the cycle strategy likewise apply information cleaning procedures to additionally refine the greater part class test (Shelke et al., 2017).

The stages of the research methodology carried out are described in the chart below :



Source : (Rahayu et al., 2020)

Gambar 1. Metodologi Penelitian

In Figure 1, we can see how the stages of the research were carried out, the dataset that was trained by the classifier in a larger training set. Before being trained, the dataset was preprocessed first, normalized to equalize the frequency from 0 to 1, numeric to nominal conversion was carried out against the class because the class was binominal, conducted resample and SMOTE sampling techniques because the data imbalance was clearly visible in the Heart Failure Clinical Records class dataset.

After preprocessing, a machine learning algorithm is used, namely the Random Forest algorithm and six other algorithms. Random forest with a classification technique that relies on "growing" a group of structured classifier trees. The proposed model is shown in Figure 1, the Heart Failure Clinical Records dataset is trained with the Random Forest algorithm and the Decision Tree algorithm, KNN, Support Vector Machine, Artificial Neural Network, Naïve Bayes in order to compare which algorithm is more suitable for use in this dataset, from each of them. - for each of these classification algorithms, the result is a correlation pattern between data that can predict the survival of heart failure patients, if it is continued it can be implemented into a prediction software.

The platform used in this study has the following specifications :

Table 3. Research Platform

Processor	intel intel® core™ i7-8565U
CPU	1.80 GHz 1.99 GHz
RAM	8.00 GB
Software	Application Weka 3.8

Source : (Rahayu et al., 2020)

The dataset that has gone through the preprocessing and classification stages is then tested using the k-fold cross validation technique, which results are seen from the accuracy, TP Rate, PCR Area, ROC Area so that it can be seen which algorithm is more suitable for this dataset as indicated by the accuracy value. Higher.

RESULT AND DISCUSSION

From the methodological stages of the research carried out, the results obtained are described in the following discussion :

1. Algorithm Testing Results

After processing, the model-trained data is generated from the results of data training after processing using the Random Forets algorithm and five other algorithms without sampling techniques. From the results of machine learning can be seen in table 1.

Table 4. Algorithm test results table

Algorithm	Accuracy	TP Rate PCR Area	ROC Area
SVM	83.61%	0.826	0.911
RF	82.60%	0.806	0.724
ANN	80.93%	0.769	0.846
DT	80.60%	0.682	0.622
NB	76.92%	0.836	0.771
KNN	68.22%	0.809	0.841

Source : (Rahayu et al., 2020)

From table 4, it can be seen that the outcomes of algorithm testing on the Failure Clinical Records dataset obtained the highest accuracy results, namely the SVM algorithm of 83.61%, RF 82.60%, ANN 80.93%, DT 80.60%, NB 76.92% and KNN 68.22%. Whereas the TP Rate PCR Area produces NB of 0.836, SVM 0.826, KNN 0.809, RF 0.806, ANN 0.769 and DT 0.682. Then for the ROC the area gets an SVM of 0.911, ANN 0.846, KNN 0.841, NB 0.771, and RF 0.724.

2. Algorithm Testing Results with SMOTE

The next test used the SMOTE sampling technique on the Failure Clinical Records dataset, which can be seen in table 5.

Table 5. Table of algorithm test results using the SMOTE technique

Algorithm	Accuracy	TP Rate PCR Area	ROC Area
RF	85.82%	0.858	0.926
SVM	80.50%	0.805	0.746
DT	80.00%	0.800	0.726
ANN	79.74%	0.797	0.869
KNN	78.22%	0.782	0.725
NB	77.21%	0.772	0.836

Source : (Rahayu et al., 2020)

From table 5, can be seen that the outcomes of algorithm testing on the Failure Clinical Records dataset obtained the highest accuracy results of RF 85.82%, SVM 80.50%, DT 80.00%, ANN 79.74%, KNN 78.22% and NB 77.21% as well as the TP Rate PCR Area which the highest is RF 0.858, SVM 0.805, DT 0.800, ANN 0.797, KNN 0.782, and NB 0.772. while for ROC the RF area is 0.926, ANN 0.869, NB 0.836, SVM 0.746, DT 0.726 and KNN 0.725.

3. Algorithm Testing Results with Resample

The next test used resample sampling technique on the Failure Clinical Records dataset using the Random Forest algorithm and five other algorithms.

Table 6. The results of algorithm testing using the resample technique

Algorithm	Accuracy	TP Rate PCR Area	ROC Area
RF	94.31 %	0.943	0.976
DT	87.29%	0.873	0.872
KNN	86.95%	0.870	0.816
SVM	79.26%	0.793	0.712
ANN	78.59%	0.786	0.860
NB	78.26%	0.783	0.850

Source : (Rahayu et al., 2020)

From table 6. can be seen that the outcomes of algorithm testing on the Failure Clinical Records dataset obtained the highest accuracy results for RF 94.31%, DT 87.29%, KNN 86.95%, SVM 79.26%, ANN 78.59% and NB 78.26% as well as for the TP Rate PCR area, namely RF of 0.943, DT 0.873, KNN 0.870, SVM 0.793, ANN 0.786 and NB 0.783. then for the ROC the area is RF of 0.976, DT 0.872, ANN 0.860, NB 0.850, KNN 0.816 and SVM 0.712.

CONCLUSION

The continuity of human life is never separated from the continuity of the body organs that work well with each other, one of the most important organs is the heart. One of the deadly heart diseases and is the final stage of heart disease is heart failure. In predicting the survival of heart failure patients, data mining techniques can be assisted. In this study, the proposed method is to use the Random Forest algorithm with preprocessing resampling techniques on the Failure Clinical Records dataset consisting of 12 attributes and 1 class which is proven to get the highest accuracy when compared to other algorithms such as Decision Tree, KNN, Support Vector Machine, Artificial Neural Network, Naïve Bayes. This test results in an accuracy of 94.31%, a TP Rate of PCR Area of 0.943 and an ROC Area of 0.976 Class, the data imbalance in the dataset used can be overcome with sample sampling techniques so that the majority of data is discarded by replacing it or without replacing it. The result is a random forest algorithm that creates a collection of decision trees from the Failure Clinical Records dataset. Further research that can be carried out by subsequent researchers can implement the patterns that are formed and build software to predict the survival of heart failure patients.

REFERECE

Aeni, W. N., Santosa, S., & Supriyanto, C. (2014). Algoritma Klasifikasi data mining naïve bayes berbasis Particle Swarm Optimization untuk deteksi penyakit jantung. *Jurnal Pseudocode*, *1*(1), 11–14. Retrieved from https://ejournal.unib.ac.id/index.php/pseudocode/article/view/57/

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients : A case study. *PLOS ONE*, 1–8.

Aninditya, A., Hasibuan, M. A., & Sutoyo, E. (2019). Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 112–117. Denpasar BALI: IEEE. https://doi.org/10.1109/IoTaIS47347.2019.8980428

Chicco, D. (2020). Heart failure clinical records Data Set. Retrieved from UCI Machine Learning Repository website: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, *20*(16), 1–16. https://doi.org/10.1186/s12911-020-1023-5

Frieden, T. R., & Berwick, D. M. (2011). The "Million Hearts" Initiative — Preventing Heart Attacks and Strokes. *The NEW ENGLAND JOURNAL of MEDICINE*, *27*(1), 1–4.

https://doi.org/10.1056/NEJMp1110421

Hamzah, R. (2016). Hubungan usia dan jenis kelamin dengan kualitas hidup pada penderita gagal jantung di RS PKU Muhammadiyah Yogyakarta (Universitas 'Aisyiyah Yogyakarta). Universitas 'Aisyiyah Yogyakarta. Retrieved from http://digilib2.unisayogya.ac.id/handle/123456789/2297

Imaligy, E. U. (2014). Gagal Jantung pada Geriatri. *CDK-212*, *41*(1), 19–24.

Manalu, E., Sianturi, F. A., & Manalu, M. R. (2017). Penerapan Algoritma Naive Bayes Untuk Memprediksi Jumlah Produksi Barang Berdasarkan Data Persediaan Dan Jumlah Pemesanan Pada Cv . Papadan Mama Pastries. *Jurnal Mantik Penusa*, *1*(2), 16–21. Retrieved from http://www.e-jurnal.pelitanusantara.ac.id/index.php/mantik/article/view/257

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *1*(1), 27–31.

Putri, S. A. (2017). INTEGRASI TEKNIK SMOTE BAGGING DENGAN INFORMATION GAIN PADA NAIVE BAYES UNTUK PREDIKSI CACAT SOFTWARE. *Jurnal Ilmu Pengetahuan Dan Teknologi Komputer*, *2*(2), 22–31.

Rahayu, S., Purnama, J. J., Pohan, A. B., Nugraha, F. S., Nurdiani, S., Hadianti, S., … Informatika, S. (2020). *Laporan Akhir Penelitian Mandiri* (Vol. 16).

Rozie, F., Hadary, F., & Wigyarianto, F. T. P. (2016). Rancang Bangun Alat Monitoring Jumlah Denyut Nadi / Jantung Berbasis Android. *Jurnal Teknik Elektro Universitas Tanjungpura*, *1*(1), 1–10. Retrieved from https://jurnal.untan.ac.id/index.php/jteuntan/article/view/13805

Sartika, D., & Sensuse, D. I. (2017). Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. *Jatisi*, *1*(2), 151–161.

Setiawati, D., Taufik, I., Jumadi, & Z, W. B. (2016). Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile. *Jurnal Online Informatika*, *I*(1), 24–27. Retrieved from http://join.if.uinsgd.ac.id/index.php/join/article/view/7

Shelke, M. S., Deshmukh, P. R., & Shandilya, P. V. K. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *International Journal of Recent Trends in Engineering & Research (IJRTER)*, *3*(4), 444–449.

Somantri, O., Wiyono, S., & Dairoh. (2016). Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM). *Scientific Journal of Informatics*, *3*(1), 34–45.

Syukri, & Samsuddin. (2018). Pengujian Algoritma Artificial Neural Network (ANN) Untuk Prediksi Kecepatan Angin. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *2*(I), 43–47. Retrieved from http://ojs.serambimekkah.ac.id/jnkti/article/view/1056

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests Estimation and Inference of Heterogeneous Treatment Effects using Random Forests ABSTRACT. *Journal Of The American Statistical Association*, *1459*. https://doi.org/10.1080/01621459.2017.1319839