

# EVALUATION OF PERFORMANCE FEATURE SELECTION ON DATA MINING TECHNIQUE FOR DIAGNOSIS CORONARY HEART DISEASE

Wiharto Wiharto<sup>1</sup>, Herianto Herianto<sup>2</sup>, Hari Kusnanto<sup>3</sup>

<sup>1</sup>Department of Informatic, Sebelas Maret University, Surakarta, Indonesia  
wiharto@staff.uns.ac.id

<sup>2</sup>Department of Mechanical and Industrial Engineering, Gadjah Mada University, Yogyakarta, Indonesia  
herianto@ugm.ac.id

<sup>3</sup>Department of Medicine, Gadjah Mada University, Yogyakarta, Indonesia  
harikusnanto@yahoo.com

## Abstract

*The process of examination of the diagnosis of coronary heart disease will generate clinical data with many attributes. The number of attributes may cause mutual contradictions between attributes, which can degrade the system performance diagnosis. This study proposes a model for a diagnosis system consisting of a combination of process, resample, removal of duplicate data, dimensional reduction and data mining techniques. Dimensional reduction techniques used are feature selection type filtering with an Information Gain (IG) algorithm. Used data mining techniques involve the classification, the algorithm k-NN, Support Vector Machine, C4.5 and multilayer perceptron (MLP). System performance is measured with a parameter sensitivity, specificity, accuracy, and Area Under the Curve (AUC). The test results indicate that the attribute generated by using information gain and the best performance is chest pain type (cp), scintigraphy (thal) and flouroscopy (ca). The performance result are for sensitivity 86.14%, specificity 89.04%, AUC 86.9% and accuracy 87.36%. The performance produced by the system is included in good categories.*

**Keywords:** Data mining, coronary heart disease, dimentional reduction, feature selection, C4.5

## 1. Pendahuluan

Perkembangan teknologi yang semakin maju berpengaruh terhadap pola hidup seseorang. Pola hidup yang tidak sehat akan berpengaruh terhadap kesehatan. Dampak pola hidup yang tidak sehat menjadi faktor pemicu terjadinya penyakit jantung koroner (Kim, Lee, & Lee, 2015), (Lee & Kim, 2016). Penyakit tersebut menurut WHO merupakan penyakit tidak menular yang menyebabkan mortalitas tertinggi di dunia. Diagnosis penyakit jantung dapat dilakukan dengan menggunakan serangkaian pemeriksaan, seperti faktor risiko, gejala, *electrocardiograph* (ECG), *scintigraphy* dan *flouroscopy*. Hasil pemeriksaan tersebut menghasilkan sejumlah data klinis dengan atribut yang relatif banyak. Pengolahan data klinis dapat dilakukan berbasis komputer dengan menggunakan teknologi *Artificial Intelligence* (AI).

Data klinis dengan jumlah yang banyak, dan disertai atribut yang banyak, akan menyebabkan informasi menjadi tersembunyi. Kondisi tersebut memerlukan sebuah teknik untuk dapat menemukan informasi tersebut. Teknik *data mining* merupakan salah satu cara yang dapat digunakan untuk mendapatkan informasi yang tersembunyi dalam sebuah kumpulan data.

Klasifikasi merupakan salah satu teknik dalam data mining. Klasifikasi mempunyai beberapa algoritma, seperti K-NN, *Multi Layer Perceptron* (MLP), C4.5 dan *Support Vector Machine* (SVM). Banyaknya atribut dalam data klinis, dapat mengakibatkan kinerja algoritma klasifikasi bekerja tidak optimal. Hal tersebut dapat saja disebabkan oleh adanya atribut yang saling kontradiktif. Untuk mengatasi hal tersebut diperlukan proses reduksi dimensi. Metode reduksi

dimensi yang dapat digunakan diantaranya *type filtering*, *wrapper* dan *embedded*. Algoritma yang digunakan dalam ketiga tipe *feature selection* tersebut, di antaranya *information gain* (IG), Algoritma genetika dan *particle swarm optimization*. Pada penelitian ini melakukan studi kinerja sistem dengan mengkombinasikan *feature selection type filtering* IG dengan algoritma klasifikasi K-NN, SVM, MLP dan C4.5. Parameter kinerja yang digunakan adalah *sensitivity*, *specificity* akurasi dan *area under the curve* (AUC), serta menggunakan validasi *10-fold cross validation*.

Banyak penelitian yang sudah dilakukan tentang penggunaan *data mining* dalam pengembangan sistem diagnosis penyakit jantung koroner. Nahar et.al (2013) dalam penelitiannya melakukan analisis kinerja sistem diagnosis dengan menggunakan beberapa algoritma klasifikasi, seperti *naive Bayesian*, SVM dan C4.5 yang dikombinasikan dengan *correlation feature selection* (CFS) dan *motivated features selection* (MFS). Pada penelitian tersebut dilakukan untuk kasus diagnosis penyakit jantung koroner dengan output *multiclass*. Kinerja dengan *10-fold cross validation* menghasilkan akurasi yang baik, sayangnya lemah dalam hal *True positive rate* (TPR) (Nahar, Imam, Tickle, & Chen, 2013).

Penelitian yang serupa dengan penelitian Nahar et.al (2013) juga dilakukan oleh Wiharto et.al (2016). Penelitian tersebut menggunakan pertimbangan *imbalanced data*, karena dataset yang digunakan terjadi *imbalanced data* pada masing-masing *classtype/level* penyakit jantung koroner. Untuk mengatasi ketidakseimbangan data dengan menggunakan SMOTE. Proses selanjutnya dilakukan *feature selection* dengan menggunakan *information gain* (IG) dan algoritma klasifikasi C4.5. Hasil pengujian dengan menggunakan *10-fold cross validation* mampu memberikan perbaikan kinerja dibandingkan tanpa mempertimbangkan menggunakan SMOTE dan IG. Selain itu dengan model tersebut mampu memperbaiki *true positive rate* yang rendah untuk *type/level* dengan *data training* yang sedikit (Wiharto, Kusnanto, & Herianto, 2016).

Penelitian berikutnya dilakukan oleh Abdar et.al (2015). Dalam penelitian tersebut dilakukan analisis kinerja sistem diagnosis penyakit jantung dengan menggunakan *logistic regression* berfungsi sebagai *feature selection*. Selanjutnya, algoritma

klasifikasi yang digunakan adalah C5.0, K-NN, SVM dan *Neural Network*. Hasilnya menunjukkan bahwa dengan menggunakan *logistic regression* mampu mereduksi atribut menjadi 6 atribut dari 13 atribut penyakit jantung koroner. Kinerja terbaik dihasilkan oleh algoritma C5.0 dengan akurasi mencapai 93,02% dan AUC 86,9%, akan tetapi penelitian tersebut tidak menggunakan validasi *K-fold cross validation*. (Abdar, Kalhori, Sutikno, Subroto, & Arji, 2015).

## 2. Data dan Metode

### 2.1. Data

Penelitian ini menggunakan dataset penyakit jantung koroner dari UCI *repository*. Pada UCI *repository* tersebut terdapat 4 dataset, salah satunya *clevelands*. Pemilihan tersebut disebabkan *dataset clevelands* tidak terdapat *missing value*. Dataset terdiri atas 14 atribut, di mana 13 atribut adalah atribut yang berpengaruh terhadap terjadinya penyakit jantung koroner, sedangkan 1 atribut adalah menandakan terjadinya penyempitan pembuluh darah *coronary*  $\geq 50\%$  atau  $\leq 50\%$ . Jumlah data dalam *dataset cleveland* berjumlah 278, dengan distribusinya 155 sehat dan 123 tidak sehat. Atribut-atribut tersebut seperti ditunjukkan pada Tabel 1.

Tabel 1. Atribut pemeriksaan yang digunakan untuk diagnosis

No	Atribut	Definisi atribut
1	Age	Age of patient
2	Sex	Gender of patient
3	Cp	Chest pain type
4	Trestbps	Resting blood pressure systolic
5	Chol	Serum cholesterol
6	Fbs	Fasting blood suger
7	Restecg	Resting ECG
8	Thalac	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST Segment
12	Ca	Number of major vessel colored by flouroscopy
13	Thal	Defect type by scintigraphy
14	Num	Diagnosis of coronary heart disease

## 2.2. Feature Selection

*Feature selection* merupakan salah satu metode dalam reduksi dimensi. Reduksi dimensi merupakan proses pengurangan dimensi dengan sedikit mungkin menghilangkan informasi (Jensen, 2005). *Feature selection* mempunyai beberapa tipe, diantaranya adalah *type filtering*. *Feature selection type filtering* melakukan proses filter dengan memberikan *score* untuk setiap atribut, selanjutnya memilih atribut-atribut berdasarkan *score* yang sudah dihitung. Contoh algoritma *feature selection type filtering* adalah *Information Gain* (IG). *Information Gain* melakukan pemberian skor berdasarkan nilai *gain* untuk masing-masing atributnya. Selanjutnya dilakukan proses perangkaian, semakin tinggi nilai *gain*-nya, maka atribut tersebut mempunyai pengaruh yang signifikan (Jain & Richariya, 2012).

## 2.3. Teknik Data Mining

*Data mining* merupakan proses menemukan pola yang menarik dari data dalam jumlah besar. *Data mining* merupakan bagian dalam proses *knowledge discovery in database* (KDD). Proses dalam KDD dapat didekomposisi menjadi 4, yaitu *data selection*, *data cleaning*, *data reduction*, *data mining* dan interpretasi (Jensen, 2005). *Data mining* mempunyai beberapa teknik, di antaranya adalah klasifikasi. Algoritma-algoritma klasifikasi dapat dikelompokkan menjadi dua yaitu *black-box* dan *non-black-box*. Algoritma *non-black-box* contohnya C4.5, ID3 dan *fuzzy inference system*. Pada penelitian ini akan dievaluasi kinerja algoritma C4.5. Algoritma tersebut merupakan pengembangan dari ID3. Algoritma C4.5 mempunyai konsep yang sama dengan ID3, tetapi jika dalam ID3 untuk menentukan atribut terpilih sebagai *root* menggunakan *information gain* tertinggi, maka dalam C4.5 menggunakan *gain ratio* (Hssina, Merbouha, Ezzikouri, & Erritali, 2014).

Algoritma klasifikasi dengan pendekatan *black-box*, seperti *Support Vector Machine* (SVM), *K-Nearest Neighborhood* (K-NN) dan *Multi Layer Perceptron* (MLP). SVM merupakan algoritma yang dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992. Kosep SVM adalah usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah class pada *input space*. Pemisah yang terbaik adalah jarak yang terjauh *hyperplane* dengan data terdekat

pada saat proses *training* (Subanya & Rajalaxmi, 2014).

Algoritma K-NN merupakan algoritma klasifikasi yang sederhana. Algoritma tersebut bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan K-NN-nya. *Training sample* diproyeksikan ke ruang berdimensi banyak, di mana masing-masing dimensi merepresentasikan atribut dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi *training sample*. Sebuah titik pada ruang ini ditandai kelas c, jika kelas c merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat dari titik tersebut. Penentuan jarak dihitung dengan menggunakan *Euclidean Distance* (Bhuvanawari & Therese, 2015).

*Multi-Layer Perceptron* (MLP) adalah jaringan syaraf tiruan *feed-forward* yang terdiri dari sejumlah *neuron* yang dihubungkan oleh bobot-bobot penghubung. *Neuron-neuron* tersebut disusun dalam lapisan-lapisan yang terdiri dari *input layer*, satu atau lebih *hidden layer*, dan *output layer*. Lapisan input menerima sinyal dari luar, kemudian melewatkannya ke lapisan tersembunyi pertama, yang akan diteruskan sehingga akhirnya mencapai lapisan output (Sonawane & Patil, 2014),(Zheng, Jiang, & Yan, 2006),(Yan, Zheng, Jiang, Peng, & Li, 2003).

## 2.4. Kinerja Sistem

Parameter kinerja yang biasa digunakan dalam dunia medis diantaranya adalah *sensitivity*, *specificity*, *area under the curve* (AUC) dan akurasi. Penentuan besarnya nilai parameter-parameter tersebut mengacu pada tabel *confusion matrix*, yang ditunjukkan pada Tabel 2. Berdasarkan tabel tersebut dapat dirumuskan perhitungan parameter kinerja tersebut seperti ditunjukkan pada persamaan (1)(2)(3).

Tabel 2. *Confusion Matrix*

Actual Class	Prediction Class	
	Positive	Negative
Positive	TP (True Positif)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (1)$$

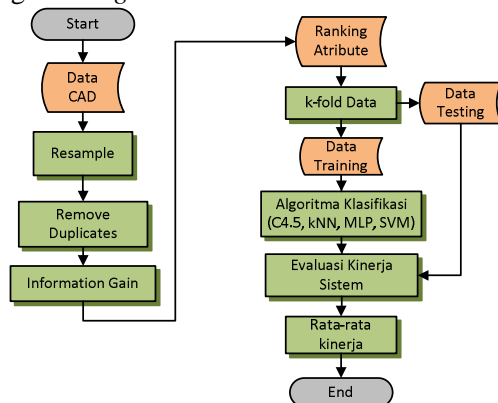
$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (2)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (3)$$

Validasi dalam pengujian sistem untuk mendapatkan parameter-parameter kinerja dilakukan dengan menggunakan metode *k-fold cross validation*. Metode tersebut membagikan data menjadi k kelompok, kemudian menggunakan k-1 kelompok data untuk *training* dan satu kelompok untuk menguji. Proses tersebut dilakukan secara bergantian, sehingga semua kelompok data pernah digunakan untuk *testing*. Nilai k berada dalam range 2...10.

### 2.5. Model Sistem yang Diusulkan

Penelitian ini melakukan evaluasi kinerja terhadap sistem diagnosis, dengan menggunakan beberapa algoritma klasifikasi, yaitu C4.5, SVM, KNN dan MLP. Model sistem yang diusulkan seperti ditunjukkan pada Gambar 1. Sebelum dilakukan proses klasifikasi dengan algoritma tersebut dilakukan beberapa proses, yaitu *resample*, *remove duplicates* dan *feature selection* dengan *information gain* (IG). Selanjutnya, dilakukan pembagian data menjadi k kelompok dengan k=10. k-1 kelompok digunakan untuk *training* dan 1 kelompok untuk *testing*. Proses tersebut dilakukan secara berulang sehingga semua kelompok pernah digunakan sebagai *testing*.



Gambar 1. Model sistem yang diusulkan

### 3. Hasil dan Analisis

Model sistem diagnosis seperti yang ditunjukkan pada Gambar 1 untuk proses *feature selection* dengan *information gain* menghasilkan atribut yang diranking ditunjukkan pada Tabel 3. Pada tabel tersebut diurutkan berdasarkan nilai *gain* dari masing-masing atribut. Atribut dengan *gain* tertinggi maka berada dalam *ranking* teratas, diikuti

dengan atribut bernilai lebih rendah. Pada penelitian ini akan dilakukan pengujian dengan menggunakan 4 model yang menunjukkan jumlah atribut yang terseleksi, seperti ditunjukkan pada Tabel 3.

Proses klasifikasi dilakukan dengan tiga skenario. Pertama dilakukan klasifikasi dengan menggunakan algoritma klasifikasi SVM, C4.5, MLP dan KNN tidak didahului dengan proses *resample* (R), *remove duplicates* (RD) dan *information gain* (IG). Kedua, dilakukan klasifikasi yang didahului proses R+RD. Ketiga, klasifikasi didahului dengan proses R+RD+IG. Skenario yang ketiga terbagi lagi menjadi 3 berdasarkan jumlah atribut yang digunakan dengan mengacu pada ranking atribut. Skenario tersebut adalah dengan menggunakan 3, 4, dan 5 atribut. Hasil pengujian dengan menggunakan *10-fold cross validation* diperoleh nilai *sensitivity* (*sens*), *specificity* (*spec*) dan akurasi(akrs), seperti yang ditunjukkan pada Tabel 4.

Tabel 3. Ranking hasil proses *Information Gain*

Ranking	Gain	Atribut	Model			
			1	2	3	4
1	0,2375	Cp	✓	✓	✓	✓
2	0,2161	ca	✓	✓	✓	✓
3	0,1982	thal	✓	✓	✓	✓
4	0,1639	oldpeak	✓	✓	✓	
5	0,1523	thalac	✓	✓		
6	0,0981	exang	✓			
7	0,0786	sex	✓			
8	0,0694	age	✓			
9	0,0598	slope	✓			
10	0,0000	chol	✓			
11	0,0000	Fbs	✓			
12	0,0000	trestbps	✓			
13	0,0000	restecg	✓			
Jumlah atribut			13	5	4	3

Kinerja sistem yang dihasilkan untuk penggunaan algoritma SVM menunjukkan bahwa dengan penggunaan model yang diusulkan yaitu R+RD+IG+SVM memberikan perbaikan kinerja dibandingkan hanya menggunakan algoritma SVM dengan 13 atribut. Selain itu, model sistem yang diusulkan juga memberikan kinerja yang lebih baik dibandingkan dengan menggunakan model R+RD+SVM dengan 13 atribut. Hal tersebut

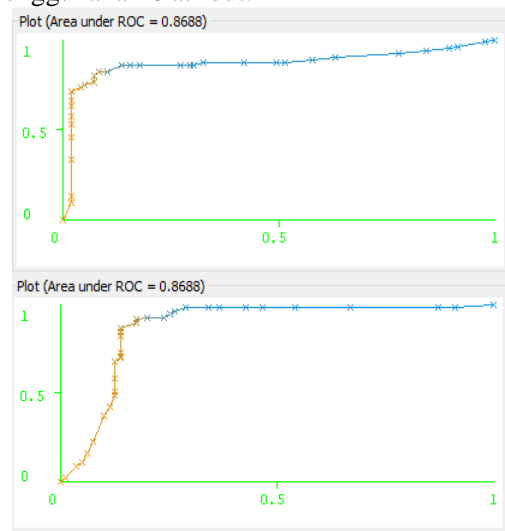
menunjukkan bahwa penggunaan *feature selection* mampu memperbaiki kinerja sistem diagnosis. Model yang diusulkan dengan menggunakan SVM memberikan kinerja terbaik, ketika menggunakan 5 atribut. Selanjutnya untuk algoritma C4.5, model sistem yang diusulkan memberikan kinerja parameter akurasi yang terbaik pada penggunaan 3 atribut. Hal tersebut sama dengan algoritma MLP dan k-NN. Jika mengacu pada kondisi terbaik untuk masing-masing algoritma, baik SVM, C4.5, MLP dan kNN, kinerja terbaik terjadi pada penggunaan algoritma C4.5. Parameter kinerja akurasi kombinasi R+RD+IG+C4.5 diperoleh 87,3%, dengan hanya menggunakan 3 atribut. Atribut tersebut adalah *chest pain type* (cp), pemeriksaan *flouroscopy* (ca) dan pemeriksaan *scintigraphy* (thal).

Tabel 4. Kinerja sistem dengan beberapa skenario

Metode	Atribut	Sens	Spec	Akrs
SVM	13	0,8385	0,8291	0,8345
R+RD+SVM	13	0,8173	0,8571	0,8333
R+RD+IG+SVM	5	<b>0,8302</b>	<b>0,8971</b>	<b>0,8563</b>
	4	0,8286	0,8841	0,8506
	3	0,8235	0,8472	0,8333
C4.5	13	0,8024	0,8108	0,8058
R+RD+C4.5	13	0,8283	0,8267	0,8276
R+RD+IG+C4.5	5	0,8763	0,8701	0,8736
	4	0,8485	0,8533	0,8506
	3	<b>0,8614</b>	<b>0,8904</b>	<b>0,8736</b>
MLP	13	0,7987	0,7647	0,7842
R+RD+MLP	13	0,8191	0,7750	0,7989
R+RD+IG+MLP	5	0,8300	0,8378	0,8333
	4	0,8400	0,8514	0,8448
	3	<b>0,8586</b>	<b>0,8667</b>	<b>0,8621</b>
kNN	13	0,8000	0,7266	0,7662
R+RD+kNN	13	0,8111	0,7381	0,7759
R+RD+IG+kNN	5	0,8000	0,7973	0,7989
	4	0,8438	0,8205	0,8333
	3	<b>0,8447</b>	<b>0,8873</b>	<b>0,8621</b>

Model sistem R+RD+IG+C4.5 dengan menggunakan 3 atribut yaitu cp, ca dan thal, jika dihitung nilai parameter kinerja AUC menghasilkan 86,9%, atau secara detail dapat ditunjukkan pada Gambar 2. Grafik AUC merupakan grafik dengan sumbu x *1-specificity* dan sumbu y *sensitivity*.

Mengacu pada nilai AUC, maka model sistem yang diusulkan dengan kombinasi R+RD+IG+C4.5, nilai AUC berada dalam rentang 80% - 90%, maka sistem diagnosis termasuk dalam kategori baik (Gorunescu, 2011). Sistem yang diusulkan jika dibandingkan dengan penelitian Abdar et.al (2015), nilai AUC bernilai sama yaitu 86,9%, tetapi untuk akurasi lebih rendah yaitu 87,36%. Perbedaan nilai akurasi tersebut disebabkan dalam penelitian Abdar et.al (2015) menggunakan model validasi yang berbeda. Penelitian Abdar et.al (2015) menggunakan model 70% data untuk *training* dan 30% untuk *testing*, sedangkan dalam sistem yang diusulkan menggunakan metode validasi yang lebih baik yaitu *10-fold cross validation*. Model validasi yang digunakan Abdar et.al (2015) belum mewakili keberagaman data yang digunakan, karena validasi hanya dilakukan satu kali. Jika pada saat pengelompokan data *training* dan *testing* distribusinya baik, maka akan menghasilkan akurasi yang baik, demikian pula sebaliknya akurasi akan rendah. Selain itu, sistem yang diusulkan hanya menggunakan 3 atribut, sedangkan pada penelitian Abdar et.al (2015) menggunakan 6 atribut.



Gambar 2. Grafik ROC untuk kombinasi RG+RD+IG+C4.5

#### 4. Kesimpulan

Model sistem diagnosis yang diusulkan dengan kombinasi *resample*, *remove duplicates*, *information gain* dan C4.5 (R+RD+IG+C4.5) mampu memberikan kinerja terbaik dibandingkan

dengan model sistem R+RD+IG yang dikombinasikan dengan SVM, MLP dan KNN. Sistem memberikan parameter kinerja *sensitivity* 86,14%, *specificity* 89,04%, akurasi 87,36% dan AUC 86,9%, dengan menggunakan *10-fold cross validation*. Kinerja tersebut hanya membutuhkan pemeriksaan *chest pain type* (cp), *flouroscopy* (ca) dan *scintigraphy* (thal). Kategori sistem dengan mengacu pada nilai AUC yang dihasilkan, maka sistem termasuk dalam kategori baik.

### Daftar Pustaka

- Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I., & Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*, 5(6), 1569–1576.
- Bhuvanawari, P., & Therese, A. B. (2015). Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm. *Procedia Materials Science*, 10, 433–440. <https://doi.org/10.1016/j.mspro.2015.06.077>
- Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques* (Vol. 12). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from <http://link.springer.com/10.1007/978-3-642-19721-5>
- Hssina, B., Merbouha, Ab., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications (IJACSA)*, (Special Issue on Advances in Vehicular Ad Hoc Networking and Applications), 13–19.
- Jain, M. M., & Richariya, V. (2012). An Improved Techniques Based on Naive Bayesian for Attack Detection. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.414.4211>
- Jensen, R. (2005). *Combining rough and fuzzy sets for feature selection* (Dissertation). University of Edinburgh. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.7352&rep=rep1&type=pdf>
- Kim, J., Lee, J., & Lee, Y. (2015). Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree. *Healthcare Informatics Research*, 21(3), 167. <https://doi.org/10.4258/hir.2015.21.3.167>
- Lee, H. K., & Kim, H. K. (2016). Lifestyle and Health-Related Quality of Life by Type-D Personality in the Patients with Ischemic Heart Disease. *Indian Journal of Science and Technology*, 9(13). <https://doi.org/10.17485/ijst/2016/v9i13/74230>
- Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, 40(1), 96–104. <https://doi.org/10.1016/j.eswa.2012.07.032>
- Sonawane, J. S., & Patil, D. R. (2014). Prediction of heart disease using multilayer perceptron neural network. In *Information Communication and Embedded Systems (ICICES)*, 2014 International Conference on (pp. 1–6). IEEE. Retrieved from [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=7033860](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7033860)
- Subanya, B., & Rajalaxmi, R. R. (2014). Feature selection using Artificial Bee Colony for cardiovascular disease classification. In *Electronics and Communication Systems (ICECS)*, 2014 International Conference on (pp. 1–6). IEEE. Retrieved from [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6892729](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6892729)
- Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Interpretation of Clinical Data Based on C4.5 Algorithm for the Diagnosis of Coronary Heart Disease. *Healthcare Informatics Research*, 22(3), 186. <https://doi.org/10.4258/hir.2016.22.3.186>
- Yan, H., Zheng, J., Jiang, Y., Peng, C., & Li, Q. (2003). Development of a decision support system for heart disease diagnosis using multilayer perceptron. In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on* (Vol. 5, p. V–709). IEEE. Retrieved from [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=1206411](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1206411)
- Zheng, J., Jiang, Y., & Yan, H. (2006). Committee machines with ensembles of multilayer perceptron for the support of diagnosis of heart diseases. In *2006 International Conference on Communications, Circuits and Systems* (Vol. 3, pp. 2046–2050). IEEE. Retrieved from [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=4064306](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4064306)

### Biodata Penulis

**Wiharto**, memperoleh gelar S1 di Sekolah Tinggi Teknologi Telkom (STT Telkom). Memperoleh gelar S2 di Universitas Gadjah Mada. Saat ini menjadi staff pengajar pada Program Studi Informatika, Universitas Sebelas Maret Surakarta.

**Herianto**, memperoleh gelar S1 di Universitas Gadjah Mada. Memperoleh gelar S2 di University of Malaya, Malaysia. Memperoleh gelar S3 dari Tokyo Institute of Technology, Jepang. Saat ini menjadi pengajar di Sekolah Pascasarjana pada Program Biomedical Engineering (BME).

**Hari Kusnanto**, memperoleh gelar S1 di Universitas Gadjah Mada, Memperoleh gelar S2 di Universitas Gadjah Mada. Memperoleh gelas S3 dari Yale University, USA. Saat ini menjadi pengajar di Fakultas Kedokteran dan Sekolah Pascasarjana pada Program Biomedical Engineering (BME) Universitas Gadjah Mada.

# BERITA ACARA PELAKSANAAN HASIL SEMINAR SESI PARALEL KNASTIK 2016

Judul : Evaluation of Performance Feature Selection on Data Mining  
Technique for Diagnosis Coronary Heart Disease

Pemakalah : Wiharto, Herianto, Hari Kusnanto

Moderator : Gloria Virginia, S.Kom., MAI, Ph.D

Notulis : Abadi

Peserta : 11 orang di ruang : D.3.2

Tanya Jawab :

Pertanyaan (oleh Gloria Virginia):

1. Data apakah yang digunakan dalam kasus penelitian tentang penyakit jantung ?
2. Apa maksud dari *re-sampling* ?

Jawaban :

1. Data klinis berupa data nominal atau ordinal seperti nilai kolesterol.
2. *Re-sampling* yaitu tahap mengambil sample data dari data yang ada kemudian tahap berikutnya adalah *me-remove* data yang *redundant*, sehingga data yang dihasilkan berkurang. Validasi ini menggunakan metode *K-Fold Cross Validation*.

Masukan Seminar :


1. Metodologi penelitian jelas dan dijelaskan dengan baik
2. Hasil perbandingan diperoleh bahwa proses *re-sampling*, *remove duplicate*, *InformationGain*, *C.A.S* merupakan komposisi yang paling baik dalam proses kategorisasi diagnosis penyakit jantung coroner.

Yogyakarta, 19 November 2016

Moderator Kelas

  
Gloria Virginia, S.Kom., MAI, Ph.D.

Penyaji Makalah

  
Wiharto