MODIFIKASI SELEKSI FITUR BERBASIS KOMPUTER UNTUK DIAGNOSIS PENYAKIT JANTUNG KORONER

Dwi Wahyu Prabowo
Email: dwi.wahyu9@gmail.com
Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Darwan Ali, Indonesia

Abstrak **- Tujuan penelitian ini adalah memodifikasi proses seleksi fitur berbasis computer/*computer based feature selection* (CFS) dengan mengganti metode *attribute selection* yang digunakan. Hal ini dilakukan untuk meningkatkan performa algoritme klasifikasi pada diagnosis penyakit jantung koroner/*coronary heart disease* (CHD) dengan jumlah atribut/fitur/faktor medis yang dipilih pada dataset Cleveland lebih sedikit. CFS II diusulkan untuk menyeleksi fitur dataset Cleveland, kemudian dilakukan pengujian untuk melihat performa yang dihasilkan. Proses seleksi fitur berdasarkan pakar medis (MFS) juga dilakukan pada penelitian ini, untuk memberikan perbandingan performa algoritme dengan CFS II. Dengan menggunakan CFS II, secara umum diperoleh peningkatan performa algoritme klasifikasi pada dataset Sick-1, Sick-2, Sick-3, dan Sick-4.**

Kata kunci **- CFS, CFS II, MFS, *attribute selection*, CHD, dataset Cleveland, dan algoritme klasifikasi.**

I. PENDAHULUAN

Coronary Heart Disease (CHD) [1] adalah contoh penyakit yang banyak diderita oleh manusia. Penyakit ini memiliki angka kematian yang sangat tinggi, contohnya pada tahun 2008 diperkirakan 7.3 juta kematian di dunia disebabkan oleh CHD [2]. CHD terjadi ketika *atherosclerosis* (timbunan lemak) menghalangi aliran darah ke otot jantung pada arteri koronaria [3].

Diagnosis awal biasanya menggunakan riwayat medis dan pemeriksaan fisik, kemudian uji lanjutan dapat dilakukan. Dari uji lanjutan ini, *coronary angiography* merupakan "standar emas" untuk mendiagnosis CHD [4]. Uji *coronary angiography* lebih dipilih oleh ahli jantung untuk mendiagnosis keberadaan CHD pada pasien dengan akurasi yang tinggi meskipun *invasive*, mempunyai risiko, dan mahal [5].

Jika dilihat dari kekurangan uji ini, perlu dikembangkan sebuah metode yang mampu mendiagnosis CHD sebelum dilakukan uji *coronary angiography*. Hal ini memotivasi pengembangan suatu metode komputer untuk dapat mendiagnosis keberadaan CHD. Metode komputer dapat menyediakan prosedur diagnosis CHD terhadap pasien dengan cara yang non-*invasive*, aman, dan lebih murah.

Banyak penelitian yang mengembangkan berbagai macam teknik komputasi cerdas untuk mendiagnosis penyakit jantung. Metode *neural network* [6][7], metode *fuzzy* [8] [9], dan *data mining* [10][11] diusulkan untuk mendiagnosis CHD. Metode *neural network* memiliki kelebihan pada prediksi *nonlinear*, kuat pada *parallel processing* dan memiliki kemampuan untuk menoleransi kesalahan, tetapi memiliki kelemahan pada perlunya data pelatihan yang besar, *over-fitting*, lambatnya konvergensi, dan sifatnya yang *local optimum* [12]. Logika *fuzzy* menawarkan penalaran pada tingkat yang lebih tinggi dengan menggunakan informasi linguistik yang diperoleh dari domain pakar, tetapi sistem *fuzzy* tidak memiliki kemampuan untuk belajar dan tidak dapat menyesuaikan diri dengan lingkungan baru [13]. *Data mining* adalah proses penggalian pengetahuan tersembunyi dari data. Metode ini dapat mengungkapkan pola dan hubungan antar sejumlah besar data dalam dataset yang tunggal atau tidak [14].

Pada diagnosis medis, reduksi data merupakan masalah penting. Data medis sering mengandung sejumlah besar fitur yang tidak relevan, *redundant*, dan sejumlah kasus yang relatif sedikit sehingga dapat mempengaruhi kualitas diagnosis penyakit [15]. Oleh karena itu, proses seleksi fitur dapat digunakan untuk menyeleksi fitur-fitur yang relevan pada data medis. Proses fitur seleksi diusulkan dalam banyak penelitian untuk meningkatkan akurasi pada proses diagnosis CHD [15] [16].

Nahar, dkk. [16] melakukan proses seleksi fitur berbasis komputer. Proses ini didefinisikan dengan *computer feature selection* (CFS). CFS memilih fitur-fitur secara random dengan cara mengalkulasi makna dari atribut-atribut data dan memperhatikan kapasitas prediksi individu. Oleh karena itu, terdapat kemungkinan untuk membuang faktor-faktor medis tertentu untuk memperoleh informasi lebih mengenai penyakit yang spesifik.

Untuk menghindari hilangnya faktor-faktor medis yang dianggap penting, perlu dilakukan proses seleksi fitur berdasarkan pakar medis/*motivated feature selection* (MFS). Faktor-faktor penting ini adalah *age, chest pain type (angina, abnang, notang, asympt), resting blood pressure, cholesterol, fasting blood sugar, resting heart rate (normal, abnormal, ventricular hypertrophy), maximum heart rate, dan exercise induced angina* [16].

Pada proses CFS, Nahar dkk. hanya menggunakan satu metode *attribute selection* saja yang disediakan oleh perangkat lunak Weka yaitu CfsSubsetEval sehingga hal ini belum mewakili secara umum mengenai proses CFS. Oleh karena itu, pada paper ini diusulkan proses seleksi fitur berbasis komputer yang menyeleksi fitur lebih sedikit namun performa algoritme klasifikasi tetap dijaga. Kemudian membandingkan hasil performa algoritme klasifikasi yang diperoleh dengan hasil pada proses MFS dan CFS.

II. DATASET DAN METODE

A. *Dataset CHD*

Pada paper ini dilakukan proses seleksi fitur pada dataset Cleveland untuk mendiagnosis CHD. Digunakan maksimum 14 atribut dari 76 atribut yang dimiliki oleh dataset Cleveland. Berikut ini dijelaskan mengenai atribut beserta tipe data yang digunakan pada dataset Cleveland [17][18].

1. *Age*: Usia dalam tahun (numerik);
2. *Sex*: *Male, female* (nominal);
3. *Chest pain type* (CP): (a*) typical angina* (*angina*), (b) *atypical angina* (*abnang*), (c) *non-anginal pain* (notang), (d) *asymptomatic* (*asympt*) (nominal). Pengertian secara medis :
   a) *Typical angina* adalah kondisi rekam medis pasien menunjukkan gejala biasa dan sehingga kemungkinan memiliki penyumbatan arteri koroner yang tinggi.
   b) *Atypical angina* adalah mengacu pada kondisi bahwa gejala pasien tidak rinci sehingga kemungkinan penyumbatan lebih rendah.
   c) *Non-anginal pain* adalah rasa sakit yang menusuk atau seperti pisau, berkepanjangan, atau kondisi menyakitkan yang dapat berlangsung dalam jangka waktu pendek atau panjang.
   d) *Asymptomatic pain* tidak menunjukkan gejala penyakit dan kemungkinan tidak akan menyebabkan atau menunjukkan gejala penyakit,
4. *Trestbps*: *resting blood pressure* pasien dalam mm Hg pada saat masuk rumah sakit (numerik);
5. *Chol*: Serum kolesterol dalam mg/dl;
6. *Fbs*: Ukuran *boolean* yang menunjukkan apakah fasting blood sugar lebih besar dari 120 mg/dl: (1 = True; 0 = false) (nominal);
7. *Restecg*: Hasil elektrokardiografi selama istirahat. Tiga jenis nilai normal (norm), abnormal (abn): memiliki kelainan gelombang ST-T, *ventricular hypertrophy* (hyp) (nominal);
8. *Thalac*: detak jantung maksimum dicapai (numerik);
9. *Exang*: Ukuran boolean yang menunjukkan apakah latihan angina induksi telah terjadi: 1 = ya, 0 = tidak ada (nominal);
10. *Oldpeak*: depresi ST yang diperoleh dari latihan relatif terhadap istirahat (numerik);
11. *Slope*: kemiringan segmen ST untuk latihan maksimum (puncak). Terdapat tiga jenis nilai yaitu condong ke atas, datar, condong ke bawah (nominal);
12. *Ca*: jumlah vessel utama (0-3) diwarnai oleh fluoroskopi (numerik);
13. *Thal*: status jantung (normal, cacat tetap, cacat reversibel) (nominal);
14. *Class*: nilai kelas baik sehat atau penyakit jantung (tipe sakit: 1, 2, 3, dan 4).

Pada paper ini, permasalahan klasifikasi *multiclass* diubah menjadi permasalahan klasifikasi biner dengan cara menganggap salah satu label kelas sebagai kelas positif dan yang lainnya dianggap sebagai kelas negatif [16]. Oleh karena itu, diperoleh 5 dataset baru yaitu H-0, Sick-1, Sick-2, Sick-3, dan Sick-4. Pada Tabel I ditunjukkan karakteristik kelima dataset.

TABEL I
KARAKTERISTIK DATASET

Nama *dataset*	**Kelas positif**	**Jumlah *instance* kelas positif**	**Jumlah *instance* kelas negatif**	**Status yang diindikasikan oleh kelas positif**
H-0	*Health*	165	138	*Health, Sick*
Sick-1	S1	54	249	S1, *Negative*
Sick-2	S2	36	267	S2, *Negative*
Sick-3	S3	35	268	S3, *Negative*
Sick-4	S4	13	290	S4, *Negative*

B. *Motivated Feature Selection (MFS)*

Motivated feature selection adalah proses seleksi fitur berdasarkan pakar medis. Terdapat delapan faktor signifikansi medis yang dipertimbangkan oleh MFS dalam proses seleksi fitur yaitu *age, chest pain type (angina, abnang, notang, asympt), resting blood pressure, cholesterol, fasting blood sugar, resting heart rate (normal, abnormal, ventricular hypertrophy), maximum heart rate, dan exercise induced angina* [16].

C. *Computer Feature Selection II (CFSII)*

CFSII merupakan metode yang diusulkan proses seleksi fitur berbasis komputer yang menyeleksi fitur lebih sedikit dibandingkan metode CFS. CFSII menggunakan *attribute selection* yang disediakan oleh perangkat lunak Weka yaitu ClassifierSubsetEval (dengan strategi pencarian BestFirst). ClassifierSubsetEval menggunakan algoritme klasifikasi sebagai parameter untuk mengevaluasi himpunan atribut pada *training* data atau pada *test* data yang terpisah [14].

D. *Algoritme Klasifikasi*

1. Naïve Bayes

Algoritme Naïve Bayes adalah *classifier probabilistic* sederhana yang berdasarkan pada penerapan teorema Bayes dengan asumsi independen yang kuat. Probabilitas *data record* X yang mempunyai label kelas C_j adalah sebagai berikut.

$$P(C_j|X) = \frac{P(X|C_j) * P(C_j)}{P(X)} \quad (1)$$

Label kelas C_j dengan nilai probabilitas bersyarat terbesar menentukan kategori data pencatatan *data record* [11].

2. SMO

Sequential Minimal Optimization (SMO) adalah algoritme untuk mengefisienkan pemecahan masalah optimasi yang muncul selama pelatihan *Support Vector Machines* (SVM). Algoritme ini diperkenalkan oleh John Platt pada tahun 1998 di Microsoft Research. SMO secara luas digunakan untuk pelatihan SVM. Terdapat dua komponen pada algoritme SMO yaitu metode analitis untuk menyelesaikan dua pengali Lagrange dan metode *heuristic* untuk menentukan pengali yang mengoptimalkan [19].

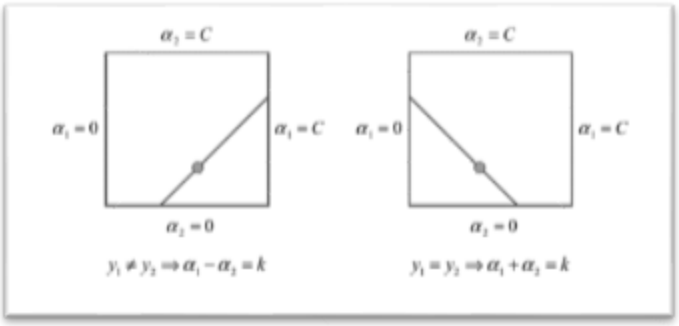
a) Solusi analitis dua pengali Lagrange

Tanpa kehilangan kondisi umum, misalkan bahwa dua elemen yang dipilih adalah α_1 dan α_2. Ketika menghitung nilai baru untuk kedua parameter, agar

tidak melanggar kendala linier $\sum_{x=1}^{l} \alpha_i y_i = 0$ maka nilai-nilai baru dari pengali harus terletak pada garis dengan formula persamaan 1 dalam ruang (α_1, α_2) dan kotak yang didefinisikan dengan $0 \leq \alpha_1, \alpha_2 \leq C$ [36].

$$\alpha_1 y_1 + \alpha_2 y_2 = konstanta = \alpha_1^{old} y_1 + \alpha_2^{old} y_2 \quad (1)$$

Gambar 1 menunjukkan garis dan kotak yang telah didefinisikan.



Gambar 1. Kedua pengali Lagrange harus memenuhi semua kendala dari masalah. Kendala ketidaksamaan menyebabkan pengali Lagrange berada di dalam kotak. Kendala kesetaraan linear menyebabkan pengali Lagrange berada pada garis diagonal. Oleh karena itu, salah satu langkah SMO harus menemukan optimum dari fungsi tujuan pada segmen garis diagonal [19].

b) Solusi analitis dua pengali Lagrange

SMO menggunakan dua kriteria untuk memilih dua titik aktif. Hal ini untuk memastikan bahwa fungsi tujuan mengalami peningkatan besar dari optimasi.

1. Titik pertama x_1 dipilih dari antara titik-titik yang melanggar kondisi Karush Kuhn Tucker [20].
2. Titik kedua x_2 harus dipilih sedemikian rupa sehingga updating pada pasangan (α_1, α_2) harus menyebabkan peningkatan yang besar pada hasil fungsi objektif ganda [20].

3. IBK

Algoritme ini menemukan sekelompok objek k dalam himpunan pelatihan yang paling dekat dengan objek uji dan dasar penugasan label pada dominasi kelas tertentu. Hal ini membahas pokok utama pada banyak dataset bahwa tidak mungkin satu objek akan persis cocok dengan objek lainnya, serta fakta bahwa informasi yang saling bertentangan tentang kelas dari sebuah objek dapat diperoleh dari objek-objek terdekat [21]. Pada Tabel II ditunjukkan algoritme k-nearest neighbour.

TABEL II
ALGORITME K-NEAREST NEIGHBOR [21]

Input	: Himpunan objek pelatihan D, Objek tes $\boldsymbol{z}$ (dalam bentuk vektor nilai-nilai atribut), dan L yaitu himpunan kelas yang digunakan untuk melabeli objek
Output	: $c_z \in D$ **do**
	Hitung $d(\boldsymbol{z}, \boldsymbol{y})$, jarak antara $\boldsymbol{z}$ dan $\boldsymbol{y}$;
End	
Pilih $N \subseteq D$, himpunan (*neighborhood*) k yang terdekat dengan objek pelatihan z;	

TABEL II
ALGORITME K-NEAREST NEIGHBOR (LANJUTAN)

$c_z = \text{argmax} \sum_{y \in N} I\left(v = kelas(c_y)\right)$;
Dengan $I(.)$ Adalah fungsi indikator yang mengembalikan nilai 1 jika argumen benar dan 0 untuk argumen salah.

4. AdaBoostM1

Tabel III menunjukkan algoritme AdaBoostM1 yang dapat digunakan untuk melakukan proses klasifikasi.

TABEL III
ALGORITME ADABOOSTM1 [14]

Model Generation

Tetapkan bobot yang sama untuk setiap contoh pelatihan.
Untuk setiap iterasi t:
- Terapkan algoritme *learning* kepada *dataset* berbobot dan simpan model yang dihasilkan.
- Hitunglah error e model pada *dataset* berbobot and simpan error.
- Jika e sama dengan nol atau e besar atau sama dengan 0,5: Hentikan model generation.
- Untuk setiap *instance* dalam *dataset*:
  - Jika *instance* diklasifikasikan dengan benar oleh model:
    - Kalikan bobot *instance* dengan $e/1-e$.
  - Normalisasi bobot semua *instance*.

Klasifikasi

Tetapkan bobot nol kepada semua kelas.
Untuk setiap (atau kurang) model t:
- Tambahkan $-\log(e/1-e)$ ke bobot kelas yang diperkirakan oleh model.

Kembali ke kelas dengan bobot tertinggi.

5. J48

J48 adalah algoritme klasifikasi yang mengimplementasikan algoritme C4.5 [14]. Algoritme C4.5 ditujukan untuk *supervised learning*. Diberikan dataset atribut bernilai dengan *instance* yang dijelaskan oleh koleksi atribut dan termasuk salah satu set kelas *mutually exclusive*, C4.5 belajar memetakan dari nilai atribut ke kelas yang dapat diterapkan untuk mengklasifikasikan kelas baru (*unseen instance*) [21]. Tabel IV menunjukkan algoritme C4.5.

TABEL IV
ALGORITME C4.5 [21]

Input: *Dataset* atribut bernilai D
1. $Tree = \{\}$
2. **if** D adalah "pure" OR kriteria pemberhentian ditemui **then**
3. terminate
4. **endif**
5. **for all** atribut $a \in D$
6. hitung kriteria *information-theoretic* jika terjadi pemisahan pada a
7. **endfor**
8. a_{best} = Atribut terbaik
9. $Tree$ = buat *node* keputusan yang menguji a_{best} pada akar
10. D_v = *Induced sub-datasets* dari D berdasarkan a_{best}
11. **for all** D_v **do**
12. $Tree_v = C4.5(D_v)$
13. *Attach* $Tree_v$ kepada cabang *Tree* yang bersesuaian
14. **endfor**
15. **return** $Tree$

6. PART

Algoritme klasifikasi PART membangun *tree* menggunakan C4.5's *heuristics* dengan parameter yang ditentukan oleh pengguna sama dengan J48. Aturan-aturan pada algoritme klasifikasi diperoleh dari *partial decision tree*. *Partial decision tree* adalah *decision tree* biasa yang mengandung cabang-cabang untuk *sub-tree* tak terdefinisi [14]. Pada Tabel V berikut ditunjukkan algoritme perluasan *partial tree*.

TABEL V
ALGORITME PART [16]

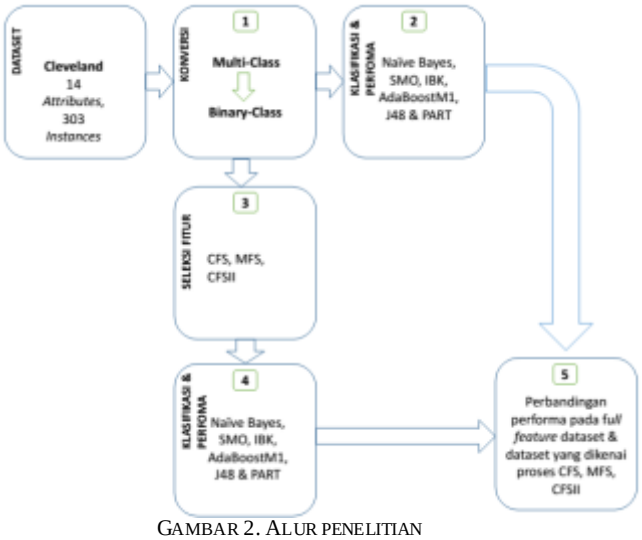
Expand-subset (S):
Pilih sebuah tes yaitu T dan gunakan untuk membagi contoh *set* menjadi *subset*
Urutkan *subset* dalam urutan menaik berdasarkan rata-rata entropi
while (terdapat *subset* X yang tidak diperluas AND semua *subset* yang diperluas adalah *leaves*)
expand-subset (X)
if (semua *subset* yang diperluas adalah *leaves* AND *error* estimasi untuk *subtree* ≥ *error* estimasi untuk *node*)
batalkan perluasan *subset* dan buat *node* sebuah *leaf*

E. Jalan Penelitian

Dataset Cleveland terlebih dahulu dikonversi dari *multi-class* menjadi *binary-class* sehingga diperoleh lima dataset dengan karakteristik sesuai dengan Tabel I. Klasifikasi pada kelima dataset ini dilakukan dengan menggunakan enam algoritme klasifikasi (Naïve Bayes, SMO, IBK, AdaBoostM1, J48, dan PART). Untuk memberikan perbandingan algoritme klasifikasi, digunakan empat matriks performa yaitu akurasi, *true positive rate* (TP), F-*measure* dan waktu pelatihan. Akurasi adalah akurasi prediksi secara keseluruhan, *true positive rate* adalah tingkat akurasi klasifikasi untuk kelas positif, dan F-*measure* adalah efektifitas algoritme ketika tingkat akurasi prediksi untuk kelas positif dan negatif diperhatikan [16].

Performa disajikan dengan melakukan *train-test split* pada dataset dan kemudian digunakan 10-*fold cross-validation* untuk memilih parameter terbaik dari algoritme klasifikasi pada proses training. Pada setiap dataset, proses *stratified sampling* digunakan untuk memilih dua pertiga data untuk pelatihan dan sisanya untuk prediksi. Salah satu alat yang disediakan oleh perangkat lunak Weka yaitu CVParameter digunakan pada proses train-test split. Kemudian dilakukan proses seleksi fitur pada kelima dataset dengan menggunakan metode MFS, CFS, dan CFSII.

Pada akhir penelitian, dilakukan perbandingan performa algoritme klasifikasi pada *full feature* dataset dengan dataset yang telah dikenai proses seleksi fitur. Untuk mendapatkan estimasi *error* yang akurat, pengulangan proses 10-*fold cross validation* sebanyak 10 kali dan kemudian hasil yang diperoleh dirata-rata merupakan prosedur standar yang harus dilakukan [14]. Oleh karena itu, pada paper ini seluruh proses yang menggunakan 10-*fold cross validation* dilakukan sebanyak 10 kali, kemudian dicari nilai rata-rata untuk hasil yang diperoleh. Gambar 2 menunjukkan alur penelitian yang dilakukan pada paper ini.



GAMBAR 2. ALUR PENELITIAN

III. HASIL PERCOBAAN

A. Seleksi Fitur

Fitur pada kelima dataset yang dipilih dengan menggunakan metode MFS, CFS, dan CFSII dapat dilihat pada Tabel VI.

TABEL VI
HASIL SELEKSI FITUR

	MFS	CFS	CFSII
H-0	Age	Chest pain	Sex
	Chest pain	Maximum heart rate	Chest pain
	Resting blood pressure	Exercise induced angina	Resting blood pressure
	Cholesterol	Oldpeak	Exercise induced angina
	Fasting blood sugar	Number of vessels coloured	Thal
	Resting ECG	Thal	
	Maximum heart rate		
	Exercise induced angina		
Sick-1	Age	Chest pain	Chest pain
	Chest pain	Exercise induced angina	Cholesterol
	Resting blood pressure	Sex	Fasting blood sugar
	Cholesterol	Thal	Oldpeak
	Fasting blood sugar	Number of vessels coloured	
	Resting ECG		
	Maximum heart rate		
	Exercise induced angina		
Sick-2	Age	Chest pain	Age
	Chest pain	Maximum heart rate	Fasting blood sugar
	Resting blood pressure	Exercise induced angina	Maximum heart rate
	Cholesterol	Oldpeak	Thal
	Fasting blood sugar	Number of vessels coloured	
	Resting ECG	Thal	
	Maximum heart rate		
	Exercise induced angina		

TABEL VI

HASIL SELEKSI FITUR (LANJUTAN)

	MFS	CFS	CFSII
Sick-3	Age	Chest pain	Sex
	Chest pain	Exercise induced angina	Chest pain
	Resting blood pressure	Fasting blood sugar	Resting blood pressure
	Cholesterol	Slope	Cholesterol
	Fasting blood sugar	Number of vessels coloured	Exercise induced angina
	Resting ECG	Thal	Thal
	Maximum heart rate		
	Exercise induced angina		
Sick-4	Age	Chest pain	Age
	Chest pain	Resting ECG	Resting ECG
	Resting blood pressure	Slope	Number of vessels coloured
	Cholesterol	Number of vessels coloured	
	Fasting blood sugar	Thal	
	Resting ECG		
	Maximum heart rate		
	Exercise induced angina		

Pada Tabel VI dapat dilihat bahwa CFSII memilih fitur yang lebih sedikit dibandingkan CFS pada dataset H-0, Sick-1, dan Sick-2. Berbeda dengan dataset Sick-3 dan Sick-4, fitur yang dipilih berjumlah sama. Jika dilihat dari kelima dataset, dapat disimpulkan bahwa fitur yang dipilih oleh CFS dan CFSII berbeda. Hal ini dikarenakan metode seleksi atribut yang digunakan berbeda.

B. Performa Algoritme Klasifikasi

Tabel VII menunjukkan performa algoritme klasifikasi yang dilakukan pada kelima dataset.

TABEL VII

PERFORMA ALGORITME KLASIFIKASI (*FULL FEATURE*)

Dataset	Algoritme	Akurasi (%)	TP	F-measure	Training Time
H-0	Naïve Bayes	**83.86137**	**0.8387**	**0.8385**	**0**
	SMO	83.26731	0.8326	0.8321	0.02
	IBK	82.07919	0.8209	0.8203	0.01
	AdaBoostM1	80.49503	0.805	0.8044	0.01
	J48	76.93067	0.7693	0.7684	**0**
	PART	78.2178	0.7822	0.7806	0.01
Sick-1	Naïve Bayes	77.2277	0.7722	0.7422	**0**
	SMO	**82.1782**	**0.822**	0.741	0.01
	IBK	81.28711	0.813	0.7405	**0**
	AdaBoostM1	79.10889	0.7911	0.7332	**0**
	J48	81.1881	0.812	**0.7499**	**0**
	PART	80.69305	0.8071	0.7358	**0**
Sick-2	Naïve Bayes	80.39602	0.804	0.8195	**0**
	SMO	**88.1188**	**0.881**	0.826	0.01
	IBK	87.92078	0.879	**0.8267**	**0**
	AdaBoostM1	85.1485	0.8513	0.8195	**0**
	J48	87.22771	0.8721	0.8228	**0**
	PART	87.32672	0.8731	0.8254	**0**
Sick-3	Naïve Bayes	83.06929	0.8306	0.8431	**0**
	SMO	**89.00989**	**0.89**	0.8386	0.01
	IBK	87.92078	0.879	0.8331	**0**
	AdaBoostM1	85.44553	0.8543	**0.8504**	**0**
	J48	88.91088	0.889	0.8381	**0**
	PART	88.1188	0.8811	0.8373	**0**
Sick-4	Naïve Bayes	94.45544	0.9445	0.9365	**0**
	SMO	**96.0396**	**0.96**	**0.941**	**0**
	IBK	95.24752	0.9521	0.9375	**0**
	AdaBoostM1	92.37623	0.9238	0.9267	**0**
	J48	**96.0396**	**0.96**	**0.941**	**0**
	PART	94.95049	0.9493	0.9355	**0**

Dataset yang digunakan pada pengujian ini adalah dataset yang belum dikenai proses seleksi fitur (*full feature*). Pada Tabel VII nilai yang diblok dengan warna hitam menunjukkan nilai performa tertinggi untuk masing-masing dataset. Algoritme klasifikasi SMO memberikan performa terbaik dalam hal akurasi dan *true positive rate* untuk dataset Sick-1, Sick-2, dan Sick-4. Pada dataset H-0, algoritme Naïve Bayes memberikan performa terbaik untuk akurasi, *true positive rate*, dan *F-Measure*. Matriks performa *training time* secara umum menunjukkan untuk semua algoritme klasifikasi pada setiap dataset, memberikan performa yang optimal. Setiap algoritme hanya membutuhkan waktu yang relatif singkat pada saat dieksekusi.

Selanjutnya perbandingan performa algoritme klasifikasi pada dataset yang telah dikenai proses seleksi fitur dapat ditunjukkan pada Tabel VIII. Matriks performa akurasi, *true positive rate*, dan *F-Measure* digunakan untuk memberikan perbandingan.

Dari hasil yang diperoleh pada Tabel VIII, seluruh performa (akurasi) algoritme klasifikasi pada proses CFSII meningkat ketika dibandingkan dengan proses MFS pada dataset H-0. Tren meningkatnya performa algoritme klasifikasi pada proses CFSII juga terjadi pada dataset Sick-1, pengecualian terjadi pada algoritme SMO (82.1782) performa akurasi sebanding antara CFSII dan MFS, serta performa akurasi algoritme IBK (80.99008) yang menurun jika dibandingkan dengan MFS.

Pada dataset Sick-2, performa akurasi algoritme PART (86.53464) pada proses CFSII menurun jika dibandingkan pada MFS. Namun, keseluruhan performa akurasi (Naïve Bayes, SMO, IBK, AdaBoostM1, dan J48) pada proses CFSII lebih unggul dibandingkan pada proses MFS. Pada dataset Sick-3 seluruh performa akurasi dari proses CFSII meningkat, pengecualian terjadi pada algoritme Naïve Bayes yang mengalami penurunan performa akurasi. Performa akurasi Naïve Bayes (84.35642) pada proses CFSII menurun jika dibandingkan dengan proses MFS.

Pada dataset Sick-4 performa akurasi algoritme SMO (96.0396) dan J48 (96.0396) sebanding untuk proses CFSII dan MFS. Kemudian algoritme Naïve Bayes (96.0396), IBK (95.84158), AdaBoostM1 (95.84158), dan PART (96.0396) lebih unggul dalam performa akurasi pada proses CFSII jika dibandingkan dengan proses MFS.

Tabel VIII juga memberikan perbandingan performa algoritme klasifikasi antara dataset yang belum dikenai seleksi fitur dengan dataset yang telah dikenai seleksi fitur. Dari hasil yang diperoleh dapat ditunjukkan bahwa pada dataset H-0, proses seleksi fitur (CFSII dan MFS) belum mampu meningkatkan performa (akurasi dan *true positive rate*) algoritme klasifikasi. Pada dataset Sick-1, proses seleksi fitur (CFSII dan MFS) mampu meningkatkan performa akurasi dan *true positive rate* untuk algoritme Naïve Bayes, AdaBoostM1, J48, dan PART. Algoritme SMO memberikan performa akurasi dan *true positive rate* yang sebanding untuk CFSII, MFS, dan *Full Feature*. Algoritme IBK memberikan performa akurasi dan *true positive rate* yang meningkat untuk proses MFS, namun menurun untuk proses CFSII.

TABEL VIII

PERFORMA ALGORITME KLASIFIKASI

Dataset	Algoritme	Akurasi (%)			TP			F-measure		
		MFS	CFSII	Full Feature	MFS	CFSII	Full Feature	MFS	CFSII	Full Feature
H-0	Naïve Bayes	75.34651	**80.69305**	**83.86137**	0.7534	**0.8069**	**0.8387**	0.7527	**0.806**	**0.8385**
	SMO	75.74255	78.71285	83.26731	0.7572	0.7872	0.8326	**0.7575**	0.7865	0.8321
	IBK	**75.84156**	78.01978	82.07919	**0.7583**	0.7802	0.8209	0.757	0.7799	0.8203
	AdaBoostM1	75.34651	77.92077	80.49503	0.7534	0.7792	0.805	0.7527	0.7781	0.8044
	J48	72.2772	76.2376	76.93067	0.7228	0.7623	0.7693	0.7219	0.7612	0.7684
	PART	74.05938	77.92077	78.2178	0.7407	0.7792	0.7822	0.7377	0.7782	0.7806
Sick-1	Naïve Bayes	81.98018	**82.1782**	77.2277	0.82	**0.822**	0.7722	0.74	**0.7491**	0.7422
	SMO	**82.1782**	**82.1782**	**82.1782**	**0.822**	**0.822**	**0.822**	0.741	0.741	0.741
	IBK	81.58414	80.99008	81.28711	0.816	0.81	0.813	0.738	0.7393	0.7405
	AdaBoostM1	82.07919	**82.1782**	79.10889	0.821	**0.822**	0.7911	0.7405	**0.7491**	0.7332
	J48	81.88117	**82.1782**	81.1881	0.819	**0.822**	0.812	**0.7425**	0.741	**0.7499**
	PART	81.78216	**82.1782**	80.69305	0.818	**0.822**	0.8071	0.7415	0.741	0.7358
Sick-2	Naïve Bayes	84.75246	87.42573	80.39602	0.8473	0.8741	0.804	0.8217	**0.8382**	0.8195
	SMO	87.82177	**88.1188**	**88.1188**	0.878	**0.881**	**0.881**	0.8258	0.826	0.826
	IBK	87.22771	87.62375	87.92078	0.8721	0.876	0.879	0.8214	0.8248	**0.8267**
	AdaBoostM1	84.25741	87.42573	85.1485	0.8426	0.8741	0.8513	0.8199	0.8251	0.8195
	J48	87.72276	87.92078	87.22771	0.877	0.879	0.8721	**0.8268**	0.8266	0.8228
	PART	**88.1188**	86.53464	87.32672	**0.881**	0.8652	0.8731	0.826	0.823	0.8254
Sick-3	Naïve Bayes	84.95048	84.35642	83.06929	0.8493	0.8434	0.8306	**0.8428**	0.8416	0.8431
	SMO	**88.91088**	**89.00989**	**89.00989**	**0.889**	**0.89**	**0.89**	0.8381	0.8386	0.8386
	IBK	88.71286	88.81187	87.92078	0.887	0.888	0.879	0.8371	0.8376	0.8331
	AdaBoostM1	86.43563	87.92078	85.44553	0.8643	0.8791	0.8543	0.8287	0.8369	**0.8504**
	J48	88.41583	**89.00989**	88.91088	0.8841	**0.89**	0.889	0.8366	**0.8437**	0.8381
	PART	87.82177	88.81187	88.1188	0.8782	0.888	0.8811	0.8409	0.8376	0.8373
Sick-4	Naïve Bayes	95.84158	**96.0396**	94.45544	0.958	**0.96**	0.9445	0.94	**0.941**	0.9365
	SMO	**96.0396**	**96.0396**	**96.0396**	**0.96**	**0.96**	**0.96**	**0.941**	**0.941**	**0.941**
	IBK	94.35643	95.84158	95.24752	0.9423	0.9581	0.9521	0.9319	0.94	0.9375
	AdaBoostM1	94.95049	95.84158	92.37623	0.9453	0.9581	0.9238	0.9334	0.94	0.9267
	J48	**96.0396**	**96.0396**	**96.0396**	**0.96**	**0.96**	**0.96**	**0.941**	**0.941**	**0.941**
	PART	95.74257	**96.0396**	94.95049	0.9571	**0.96**	0.9493	0.9395	**0.941**	0.9355

Pada dataset Sick-2, proses seleksi fitur (CFSII dan MFS) mampu meningkatkan performa akurasi dan *true positive rate* untuk algoritme Naïve Bayes dan J48. Algoritme IBK memberikan penurunan performa akurasi dan *true positive rate* ketika dilakukan proses CFSII dan MFS. Algoritme SMO memberikan penurunan performa akurasi dan *true positive rate* ketika dilakukan proses MFS, namun memberikan performa akurasi dan *true positive rate* yang sebanding ketika dilakukan proses CFSII. Algoritme AdaBoostM1 memberikan penurunan performa akurasi dan *true positive rate* ketika dilakukan proses MFS, namun memberikan peningkatan performa akurasi dan *true positive rate* ketika dilakukan proses CFSII. Algoritme PART memberikan peningkatan performa akurasi dan *true positive* ketika dilakukan proses MFS, namun memberikan penurunan performa akurasi dan *true positive* ketika dilakukan proses CFSII.

Pada dataset Sick-3, proses seleksi fitur (CFSII dan MFS) mampu meningkatkan performa akurasi dan *true positive rate* untuk algoritme Naïve Bayes, IBK, dan AdaBoostM1. Algoritme SMO memberikan penurunan performa akurasi dan *true positive rate* ketika dilakukan proses MFS, namun memberikan performa akurasi dan *true positive rate* yang sebanding ketika dilakukan proses CFSII. Algoritme J48 dan PART memberikan penurunan performa akurasi dan *true positive rate* ketika dilakukan proses MFS, namun memberikan peningkatan performa akurasi dan *true positive rate* ketika dilakukan proses CFSII.

Pada dataset Sick-4, proses seleksi fitur (CFSII dan MFS) mampu meningkatkan performa akurasi dan *true positive rate* untuk algoritme Naïve Bayes, AdaBoostM1, dan PART. Algoritme IBK memberikan penurunan performa akurasi dan *true positive rate* ketika dilakukan proses MFS, namun memberikan peningkatan performa akurasi dan *true positive rate* ketika dilakukan proses CFSII. Algoritme SMO dan J48 memberikan performa akurasi dan *true positive rate* yang sebanding ketika dilakukan proses CFSII dan MFS.

IV. KESIMPULAN

Jika ditinjau ulang dari Tabel VII dimulai dari dataset Sick-1 hingga Sick-4, proses seleksi fitur CFSII cenderung mampu meningkatkan performa akurasi dan *true positive rate*. Oleh karena itu, proses seleksi fitur pada dataset Cleveland mampu memberikan peluang untuk meningkatkan performa diagnosis CHD. Namun jika mengandalkan proses seleksi fitur berbasis komputer saja, hal ini merupakan sebuah kesalahan. Kebutuhan mengenai terlibatnya para ahli/dokter untuk mendiagnosis CHD harus dipenuhi.

Berdasarkan Tabel VI, tidak ada yang bisa menjamin bahwa dokter akan percaya mengenai fitur/faktor medis yang dipilih secara otomatis oleh komputer. Oleh karena itu untuk penelitian selanjutnya, dapat diusulkan penggabungan metode CFSII dan MFS sebagai bentuk keterlibatan para ahli/dokter dalam usaha mendiagnosis CHD melalui bantuan komputer, sehingga tingkat kepercayaan dokter akan hasil diagnosis yang diperoleh meningkat. Modifikasi terhadap metode *attribute selection* pada proses CFSII juga dapat dilakukan, hal ini bertujuan untuk memperoleh peningkatan performa algoritme klasifikasi pada seluruh dataset (H-0, Sick-1, Sick-2, Sick-3, dan Sick-4).

REFERENSI

[1] A. Selzer, Understanding Heart Disease. University of California Press, 1992.

[2] WHO, Global Atlas on Cardiovascular Disease Prevention and Control, 1st ed. World Health Organization, 2012.

[3] O. S. Randall, N. M. Segerson, and D. S. Romaine, The Encyclopedia of the Heart and Heart Disease, 2nd ed. Facts on File, 2010.

[4] B. Phibbs, The Human Heart: A Basic Guide to Heart Disease, Second. Lippincott Williams & Wilkins, 2007.

[5] N. A. Setiawan, "Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System," Universiti Teknologi Petronas, 2009.

[6] A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," presented at the 2011 21st International Conference on Systems Engineering (ICSEng), 2011, pp. 406–409.

[7] K. Rajeswari, V. Vaithiyanathan, and T. R. Neelakantan, "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks," Procedia Eng., vol. 41, pp. 1818–1823, 2012.

[8] V. Khatibi and G. A. Montazer, "A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment," Expert Syst. Appl., vol. 37, no. 12, pp. 8536–8542, 2010.

[9] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," J. King Saud Univ. - Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 2012.

[10] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," presented at the 2012 Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), 2012, pp. 173–177.

[11] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease," Comput. Methods Programs Biomed., 2013.

[12] R. Capparuccia, R. De Leone, and E. Marchitto, "Integrating support vector machines and neural networks," Neural Netw., vol. 20, no. 5, pp. 590–597, Jul. 2007.

[13] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems, 2nd ed. Addison-Wesley, 2004.

[14] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, 2nd ed. Morgan Kaufmann, 2005.

[15] N. Chu, L. Ma, J. Li, P. Liu, and Y. Zhou, "Rough set based feature selection for improved differentiation of traditional Chinese medical data," presented at the 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2010, vol. 6, pp. 2667–2672.

[16] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," Expert Syst. Appl., vol. 40, no. 1, pp. 96–104, Jan. 2013.

[17] UCI, "Heart disease dataset," 28-May-2013. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleve.mod. [Accessed: 28-May-2013].

[18] UCI, "Cleveland heart disease data details," 19-Jun-2013. [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names. [Accessed: 19-Jun-2013].

[19] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

[20] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines: and other kernel-based learning methods. Cambridge; New York: Cambridge University Press, 2000.

[21] X. Wu and V. Kumar, The top ten algorithms in data mining. Boca Raton: CRC Press, 2009.