JURNAL RESPONSIF, Vol. 2 No.2 Agustus 2020, pp. 213~221
E-ISSN**:** 2685-6964

SELEKSI FITUR DAN OPTIMASI PARAMETER *k-NN* BERBASIS ALGORITMA GENETIKA PADA DATASET MEDIS

Rizki Tri Prasetio

Universitas Adhirajasa Reswara Sanjaya
e-mail: rizki@ars.ac.id

Abstrak

Klasifikasi dataset medis adalah masalah data mining utama yang sedang diteliti selama satu dekade yang telah menarik beberapa peneliti dari berbagai bidang. Banyak algoritma klasifikasi dirancang untuk belajar dari data itu sendiri melalui proses pelatihan, karena pengetahuan ahli untuk menentukan parameter klasifikasi sulit. Penelitian ini mengusulkan metodologi yang didasarkan pada paradigma data mining. Paradigma ini mengintegrasikan pencarian heuristic yang terinspirasi dari evolusi alam yang disebut algoritma genetika dengan algoritma pembelajaran yang paling sederhana dan paling banyak digunakan, k nearest neighbor. Algoritma genetika digunakan untuk pemilihan fitur dan optimasi parameter sedangkan k-nearest neighbor digunakan sebagai algoritma klasifikasi. Metode yang diusulkan diujicobakan pada lima dataset medis dari UCI Machine Learning Repository untuk menangani klasifikasi dataset medis. Hasil percobaan menunjukkan bahwa metode yang diusulkan mampu mencapai kinerja yang baik, dibandingkan dengan hasil pengklasifikasi lain dengan peningkatan yang signifikan dengan nilai p uji-t 0.0011.

Kata Kunci*:* klasifikasi, algoritma genetika, fitur seleksi, optimasi parameter, *k-nearest neighbor*

Abstract

Medical dataset classification is a major data mining problem being researched about for a decade that has attracted several researchers from different fields. Most classifiers are designed to learn from the data itself through training process, because expert knowledge to determine classifier parameters is difficult. This research proposes a methodology based on data mining paradigm. This paradigm integrates the search heuristic that is inspired by natural evolution called genetic algorithm with the simplest and the most used learning algorithm, k-nearest Neighbors. The genetic algorithm is used for feature selection and parameter optimization while k-nearest Neighbors is used as a classifier. The proposed method is experimented on five medical datasets of the UCI Machine Learning Repository for handling medical dataset classification. Experiment results show that the proposed method is able to achieve good performance, compared to the results of other classifiers with significant improvement with p value of t-Test is 0.0011.

Keywords: *classification, genetic algorithm, features selection, parameters optimization, k-nearest neighbor*

Naskah diterima 26 Juli 2020; direvisi 2 Agustus 2020; diterbitkan 31 Agustus 2020

1. Pendahuluan

Penerapan pembelajaran mesin dalam diagnosis medis menjadi tren utama untuk aplikasi data medis baru-baru ini. Sebagian besar teknik diagnosis di bidang medis disistematisasi sebagai pendekatan klasifikasi data cerdas (Subbulakhsmi & Deepa, 2015). Penggunaan sistem *Computer-Aided Diagnosis* (CAD) dapat membantu dokter untuk mendiagnosis penyakit pasien (Unal & Kocer, 2013), di antara berbagai tugas yang dilakukan oleh sistem CAD, klasifikasi adalah yang paling umum. Masalah klasifikasi dataset medis dapat dikategorikan sebagai kelas masalah optimasi yang kompleks dengan tujuan untuk menjamin bantuan diagnosis secara akurat (Subbulakhsmi & Deepa, 2015).

Berbagai peneliti telah mencoba menerapkan berbagai teknik untuk meningkatkan akurasi klasifikasi data untuk mengidentifikasi pasien potensial (Babu & Suresh, 2013). Dalam studi terbaru, algoritma metaheuristik seperti *particle swarm optimization* (Subbulakhsmi & Deepa, 2015) (Inbarani, et al., 2014) (Chang, et al., 2012) atau algoritma genetika (Raymer, et al., 2000) (Yang & Honavar, 1998) (Shah & Kusiak, 2007) dan juga teknik data mining lain seperti *neural networks* (Mazurowski, et al., 2008) (Brameier & Banzhaf, 2001) (Amato, et al., 2013) atau *k-nearest neighbor* (Prasetio & Pratiwi, 2015) (Suguna & Thanushkodi, 2010) (Jabbar, et al., 2013) diterapkan untuk klasifikasi data medis dan memperoleh hasil yang sangat baik.

Algoritma *k-nearest neighbours* (k-NN) merupakan salah satu metode yang menggunakan *supervised algorithm* (Wu, et al., 2008). Dimana k-NN merupakan teknik klasifikasi yang mudah dipahami dan diterapkan (Wu & Kumar, 2009) dan paling sederhana di antara semua algoritma pembelajaran mesin (Gorunescu, 2011). Metode k-NN merepresentasikan teknik untuk mengklasifikasikan suatu objek berdasarkan objek terdekat (k) di sekitarnya (Harrington, 2012). k-NN sangat cocok untuk kelas multimodal serta aplikasi di mana suatu objek dapat memiliki banyak label kelas (Wu & Kumar, 2009).

Ada beberapa masalah utama yang memengaruhi performa k-NN. Salah satunya adalah pemilihan parameter *k* (Wu & Kumar, 2009). Jika *k* terlalu kecil, maka hasilnya dapat sensitif terhadap titik-titik noise yang dapat mengarahkan algoritma pada kondisi *overfitting* (Larose, 2005). Di sisi lain, jika *k* terlalu besar, lingkungan tersebut mungkin menyertakan terlalu banyak poin dari kelas lain (Wu, et al., 2008). Pilihan parameter terbaik dari *k* bergantung pada data (Gorunescu, 2011).

Masalah utama lainnya adalah akurasi algoritma k-NN dapat sangat terdegradasi dengan adanya fitur yang *noise* atau tidak relevan. (Han, et al., 2012), atau jika skala fitur tidak konsisten dengan kepentingannya (Gorunescu, 2011).

Data medis yang terlibat dalam model diagnostik biasanya berdimensi tinggi. Kumpulan data berdimensi tinggi meningkatkan kompleksitas klasifikasi dan mengurangi efek model (Bharti & Singh, 2014), hambatan serius bagi efisiensi sebagian besar algoritma data mining. Rintangan ini terkadang dikenal sebagai "kutukan dimensi" atau "*curse of dimensionality*" (Maimon & Rokach, 2010).

Dimensi data perlu dikurangi namun tetap mempertahankan informasi penting. Ekstraksi fitur (Liu, et al., 2015) dan fitur seleksi (Jirapech-Umpai & Aitken, 2005) adalah metode utama dalam reduksi dimensi. Proses data mining membutuhkan biaya komputasi yang tinggi saat menangani kumpulan data yang besar. Mengurangi dimensi dapat secara efektif memotong biaya (Maimon & Rokach, 2010), mengurangi waktu eksekusi dan penggunaan kapasitas memori (Shilaskar & Ghatol, 2013).

Tujuan utama dari pemilihan fitur adalah untuk mengurangi jumlah fitur yang digunakan dalam klasifikasi dengan tetap menjaga akurasi klasifikasi yang dapat diterima (Raymer, et al., 2000). Pemilihan fitur dapat berdampak besar pada keefektifan algoritma klasifikasi yang dihasilkan (Jain & Zongker, 1997), dalam beberapa kasus, sebagai hasil dari pemilihan fitur, akurasi klasifikasi yang akan datang dapat ditingkatkan (Maimon & Rokach, 2010).

Masalah pemilihan fitur didefinisikan sebagai sekumpulan fitur kandidat dan mengoptimalkan subset yang berkinerja terbaik di bawah beberapa sistem klasifikasi (Jain & Zongker, 1997). Algoritma genetika sering digunakan untuk melakukan optimasi. Algoritma genetika memiliki kecenderungan yang lebih kecil untuk terjebak dalam minimum local atau *local minima*

(Gorunescu, 2011), dan pengoptimalan yang tergolong lebih canggih (Witten, et al., 2011).

Algoritma genetika adalah berbagai macam pengoptimalan yang bergantung pada fungsi tujuan (fitness) (Prasetio & Riana, 2015), mudah diparalelkan dan telah digunakan untuk klasifikasi serta masalah pengoptimalan lainnya. Dalam data mining, mereka dapat digunakan untuk mengevaluasi kesesuaian algoritma lain (Han, et al., 2012).

Dalam penelitian ini, diintegrasikan algoritma genetika untuk pemilihan fitur dan parameter yang dioptimalkan pada k-NN yang diujicobakan untuk mengklasifikasikan lima dataset medis yang dijadikan tolak ukur berbagai penelitian sebelumnya, yaitu *breast-cancer prognostic and diagnostic* Wisconsin. (Mangasarian, et al., 1995), *diabetic retinopathy* Debrecen (Antal & Hajdu, 2014), *cardiotocography* (Ayres-de-campos, et al., 2000) dan *SPECTF image of heart disease* (Kurgan, et al., 2001).

Tabel 1. Deskripsi Dataset

Dataset	Jumlah Sampel	Jumlah Fitur	Jumlah Kelas
Wisconsin Breast Cancer (Diagnostic)	569	32	2
Wisconsin Breast Cancer (Prognostic)	198	34	2
Diabetic Retinopathy Debrecen	1151	20	2
Cardiotocography (CTGs)	2126	23	3
Heart Disease (SPECTF)	267	44	2

Tujuan utama dari penelitian ini adalah untuk meningkatkan akurasi klasifikasi lima medical dataset yang umum digunakan dengan menerapkan algoritma genetika sebagai fitur seleksi dan meningkatkan performansi algoritma k-NN dengan mengoptimalkan nilai *k* menggunakan algoritma genetika.

2. Metode Penelitian

2.1. Dataset

Penelitian ini diujicobakan pada lima dataset medis yang diperoleh dari *UCI Machine Learning Repository*. Rincian dataset medis yang digunakan dapat dilihat pada Tabel 1 yang berisi jumlah sampel, fitur, dan kelas. Dataset *training* dan *testing* dibuat secara acak.

Wisconsin *Breast Cancer* (*Diagnostic*), merupakan kumpulan data yang tersedia di University of Wisconsin. Data ini berisi 569 kasus dengan 32 fitur yang digunakan untuk memprediksi pertumbuhan kanker payudara jinak atau ganas (Mangasarian, et al., 1995).

Wisconsin *Breast Cancer* (Prognostic), dataset diperoleh dari University of Wisconsin. Terdapat 198 kasus dengan 20 fitur yang digunakan untuk memprediksi kemungkinan kanker payudara tersebut dapat berulang dan tidak berulang (Mangasarian, et al., 1995).

Diabetic Retinopathy, dataset ini dikumpulkan dari University of Debrecen dan berisi sekitar 1151 kasus dengan 20 fitur yang digunakan untuk memprediksi apakah pasien mengidap diabetes retinopati atau tidak. (Antal & Hajdu, 2014).

Cardiotocography (CTGs), merupakan kumpulan data yang dibuat oleh Diogo Ayres-de-campos di Universitas Porto. Data ini berisi 2126 kasus dengan 23 fitur yang digunakan untuk memprediksi keadaan janin (Ayres-de-campos, et al., 2000).

Heart Disease (SPECTF), kumpulan data ini didasarkan pada data dari University of Colorado. Data ini berisi 45 fitur dengan 267 kasus yang digunakan untuk mengidentifikasi apakah pasien normal atau tidak (Kurgan, et al., 2001).

2.2. Algoritma Genetika

Algoritma genetika (GA) adalah algoritma pencarian stokastik, paralel, heuristik yang terinspirasi oleh prinsip dasar seleksi alam yang diperkenalkan oleh Charles Darwin (Nowe, 2014). Prinsip dasar GA pertama kali dikemukakan oleh Holland (Holland, 1975).

Algoritma genetika mencoba meniru proses komputasi dimana seleksi alam merupakan proses biologis di mana individu yang lebih kuat kemungkinan besar menjadi pemenang dalam lingkungan yang bersaing. (Man, et al., 1996) mengoperasikan dan menerapkannya untuk memecahkan masalah bisnis dan penelitian (Larose, 2006).

Algoritma genetika menyediakan kerangka kerja untuk mempelajari efek dari faktor-faktor yang diilhami secara biologis seperti pemilihan pasangan, reproduksi, mutasi, dan persilangan informasi genetik. Tiga operator digunakan oleh algoritma genetika:(Gorunescu, 2011)

1. *Selection*. Operator seleksi mengacu pada metode yang digunakan untuk memilih kromosom mana yang akan bereproduksi. Fungsi kesesuaian mengevaluasi setiap kromosom (larutan kandidat), dan semakin baik kromosom tersebut, semakin besar kemungkinannya akan dipilih untuk bereproduksi.
2. *Crossover*. Operator *crossover* melakukan rekombinasi, menciptakan dua keturunan baru dengan memilih lokus secara acak dan bertukar urutan ke kiri dan kanan lokus tersebut antara dua kromosom yang dipilih selama seleksi. Misalnya, dalam representasi biner, dua string 11111111 dan 00000000 dapat disilangkan di lokus keenam di masing-masing untuk menghasilkan dua keturunan baru 11111000 dan 00000111.

Algoritma 1. Algoritma Genetika

Begin
INITIALIZE inisialisasi populasi dengan kandidat solusi yang acak;
EVALUATE evaluasi setiap kandidat;
REPEAT UNTIL (kondisi penghentian terpenuhi) *DO*
1. *SELECT* pilih induk;
2. *RECOMBINE* pasangkan induk;
3. *MUTATE* mutasi keturunan yang dihasilkan;
4. *EVALUATE* evaluasi kandidat baru;
5. *SELECT* pilih individu untuk generasi berikutnya;

End

3. *Mutation*. Operator mutasi secara acak mengubah bit atau digit pada lokus tertentu dalam kromosom: biasanya, bagaimanapun, dengan kemungkinan yang sangat kecil. Misalnya, setelah persilangan, string anak 11111000 dapat bermutasi pada lokus dua menjadi 10111000. Mutasi memperkenalkan informasi baru ke kumpulan genetik dan melindungi agar tidak terlalu cepat berkumpul ke optimal lokal.

2.3. *k-Nearest Neighbor*

Algoritma *k-nearest neighbor* (k-NN) merupakan salah satu metode yang menggunakan *supervised algorithm* (Wu, Kumar, Quinlan, Ghosh, & Yang, 2008). Dimana merupakan algoritma yang paling sederhana (Gorunescu, 2011), sering digunakan untuk klasifikasi, meskipun dapat juga digunakan untuk estimasi dan prediksi.

k-nearest neighbor adalah contoh pembelajaran berbasis sampel, di mana kumpulan data pelatihan disimpan, sehingga klasifikasi untuk catatan baru yang tidak diklasifikasikan dapat ditemukan hanya dengan membandingkannya dengan catatan yang paling mirip di set pelatihan (Larose, Data Mining Methods and Models, 2006).

Algoritma k-NN merepresentasikan teknik untuk mengklasifikasikan suatu objek berdasarkan objek terdekat (*k*) di sekitarnya (Gorunescu, 2011) dan mendasarkan penetapan label pada dominasi kelas tertentu di lingkungan terdekat ini (Wu & Kumar, The Top Ten Algorithms in Data Mining, 2009). Kami melihat pada *k* bagian data yang paling mirip dari kumpulan data kami yang diketahui; dari sinilah *k* berasal (Harrington, 2012).

Algoritma 2. *k-Nearest Neighbor*

Input:
D, set data training, set data testing,
z, vektor nilai atribut,
L, set kelas yang digunakan untuk memberi label pada objek

Output: $c_z \in L$, the class of **z**
foreach object **y** $\in D$ do
| Hitung d(**z, y**), jarak antara **z** dan **y**;
end

Select $\boldsymbol{N} \subseteq D$, himpunan (tetangga terdekat) dari *k* objek pelatihan terdekat ke **z**;
$c_z = \text{argmax} \sum_{y \in N} I(v = \text{class}(c_y))$;

dimana *I* adalah fungsi indikator yang mengembalikan nilai 1 jika argumennya benar dan 0 sebaliknya

Untuk menyimpulkan, algoritma *k-nearest neighbor* adalah salah satu yang paling sederhana dari semua algoritma pembelajaran mesin, karena algoritma ini hanya terdiri dari pengklasifikasian objek dengan suara mayoritas tetangganya. (Gorunescu, 2011) dari *k* bagian data yang paling mirip (Harrington, 2012).

Kedekatan antara objek dengan lingkungannya didefinisikan dalam metrik jarak, seperti *Euclidean Distance* atau *Manhattan Distance* (Han, Kamber, & Pei, 2012). Untuk membangun algoritma, kita

membutuhkan beberapa persiapan sebagai berikut:

1. Satu set data yang memiliki label (Gorunescu, 2011) yang akan digunakan untuk mengevaluasi kelas objek uji (Wu, Kumar, Quinlan, Ghosh, & Yang, 2008) (*training dataset*);
2. Jarak (metrik) untuk menghitung kemiripan antar objek (Gorunescu, 2011) yang dapat digunakan untuk menghitung kedekatan objek (Wu, Kumar, Quinlan, Ghosh, & Yang, 2008);
3. Nilai *k*, banyaknya tetangga terdekat (Wu, Kumar, Quinlan, Ghosh, & Yang, 2008) milik set data pelatihan, berdasarkan klasifikasi objek baru yang akan kita capai (Gorunescu, 2011);
4. Metode yang digunakan untuk menentukan kelas objek sasaran berdasarkan kelas dan jarak k tetangga terdekat (Wu & Kumar, 2009)

Berdasarkan empat kebutuhan yang perlu dipersiapkan, objek baru (belum diklasifikasikan) akan diklasifikasikan dengan melakukan langkah-langkah berikut: (Gorunescu, 2011)

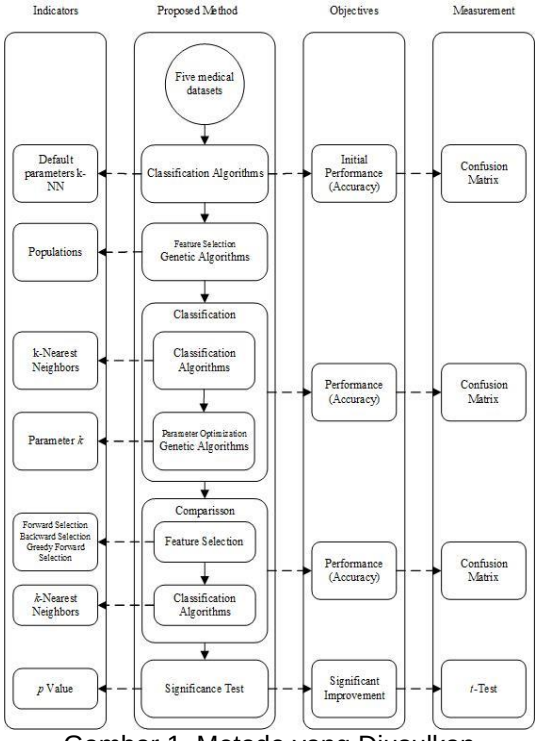
1. Hitung jarak (kesamaan) antara semua data latih dan objek baru (pendekatan naif melalui penghitungan jarak);
2. Identifikasi *k* objek terdekat (*k* tetangga paling mirip), dengan mengurutkan objek pelatihan dengan memperhitungkan jarak yang dihitung pada langkah pertama;
3. Tetapkan label yang paling sering di antara *k* pada data latih yang terdekat dengan objek itu (pemungutan suara mayoritas atau *majority rule*).

2.4. Metode yang Diusulkan

Metode yang diusulkan mengintegrasikan algoritma genetika untuk pemilihan fitur dan parameter yang dioptimalkan k-NN berlaku untuk mengklasifikasikan lima dataset medis yang dijadikan patokan dijelaskan pada Tabel 1. Metode yang diusulkan dapat dilihat pada Gambar 1. Pengolahan data awal dimulai dengan membagi lima dataset menjadi pelatihan dan menguji data menggunakan validasi terpisah. k-NN dengan parameter default diterapkan untuk setiap data pelatihan untuk menghasilkan performa awal.

Algoritma genetika diterapkan untuk setiap data pelatihan untuk pemilihan fitur. Pilihan fitur digunakan untuk menemukan fitur yang paling mewakili kelas pada dataset tersebut. Parameter yang dioptimalkan k-NN kemudian diterapkan pada data latih yang telah dipilih fitur. Setelah itu dilakukan validasi model yang dihasilkan oleh k-NN, hitung seberapa besar akurasi yang dihasilkan oleh model yang diujikan pada data pengujian. Jika akurasi yang diinginkan belum tercapai, ulangi proses pemilihan fitur menggunakan algoritma genetika. Iterasi ini akan terus dilakukan hingga didapatkan fitur yang optimal.

Hasil yang diperoleh dari performansi seleksi fitur oleh algoritma genetika kemudian dibandingkan dengan algoritma lain yang dapat digunakan untuk seleksi fitur yaitu *backward elimination*. (Guyon & Elisseeff, 2003) (Abe, 2005) (Derksen & Keselman, 1992), *forward selection* (Blanchet, et al., 2008) (Abe, 2010) (Jain & Zongker, 1997) dan *greedy feature selection* (Dyer, et al., 2013) (Vafaie & Imam, 1994) (Farahat, et al., 2013). Perbandingan ini untuk menentukan apakah performa dari algoritma genetika lebih baik dari algoritma lainnya dalam melakukan seleksi fitur.



Gambar 1. Metode yang Diusulkan

Hasil yang diperoleh dari metode yang diusulkan kemudian diuji dengan hasil yang diperoleh dari k-NN dengan parameter default untuk mengetahui apakah hasil kinerja metode yang diusulkan meningkatkan akurasi kelima dataset secara

signifikan menggunakan uji-t (Prasetio & Pratiwi, 2015) (Setiyorini & Wahono, 2015) (Prasetio & Riana, 2015).

3. Hasil dan Pembahasan

Pada penelitian ini dilakukan beberapa percobaan, percobaan menggunakan algoritma k-NN dengan parameter yang belum dioptimasi dari lima dataset yang tidak dilakukan seleksi fitur, percobaan menggunakan algoritma k-NN dengan parameter yang dioptimalkan dari lima dataset pada Tabel 1 yang belum terseleksi fitur dan percobaan menggunakan Algoritma k-NN dengan parameter optimal dari lima dataset pada Tabel 1 yang telah diseleksi fiturnya menggunakan algoritma genetika, *backward elimination*, *forward selection* dan *greedy feature selection*.

Semua eksperimen menggunakan validasi *split validation* untuk memisahkan kumpulan data secara acak. Percobaan menggunakan konfigurasi parameter default untuk algoritma genetika, *backward elimination*, *forward selection* dan *greedy feature selection*.

Tabel 2. Hasil Eksperimen dari Metode yang Diusulkan

Dataset	Metode yang Diusulkan	*k*-NN
breast-cancer (D)	99.2.%	94.15%
breast-cancer (P)	86.44%	78.75%
diabetic-retinopathy	71.69%	61.16%
cardiotocography	98.59%	90.91%
heart (SPECTF)	87.5%	77.5%

Hasil eksperimen yang dituangkan dalam Tabel 2 menyatakan bahwa metode yang diusulkan dapat meningkatkan akurasi kelima dataset benchmark dengan peningkatan 5% - 10% dibandingkan dengan algoritma k-NN tanpa optimasi dan pemilihan fitur.

Peningkatan kinerja tertinggi diperoleh dari klasifikasi dataset *diabetic retinopathy* dengan peningkatan 10,53% sebesar 61,16% dengan *k* paling optimal 6. Sedangkan peningkatan kinerja terendah diperoleh dari klasifikasi dataset *breast-cancer diagnostic* dengan hanya 5,05% meningkat dari 94,15% dengan *k* optimal 8.

Peningkatan performansi pada dataset *breast-cancer prognostic* sebesar 7,69% dari 78,75% dengan *k* optimal 5, dataset *cardiotocography* meningkat 7,68% dari semula 90,91% dengan *k* optimal 3 dan dataset SPECTF *heart disease* meningkat 10% dari 77,5% dengan yang paling optimal *k* adalah 3.

Berdasarkan hasil percobaan dalam penelitian ini, untuk mengetahui apakah metode yang diusulkan dapat meningkatkan kinerja dalam klasifikasi dataset medis secara signifikan. Pengujian menggunakan uji signifikansi dilakukan, *t-Test Paired Two Sample for Means* yang digunakan dalam hasil antara sebelum dan sesudah menggunakan metode yang diusulkan.

Tabel 3. Hasil Eksperimen dari Seleksi Fitur

Dataset	Proposed Method	Forward Selection	Backward Selection	Greedy Selection
breast-cancer (D)	99.2%	98.83%	97.08%	92.4%
breast-cancer (P)	86.44%	84.75%	83.05%	79.66%
Diabetic-retinopathy	71.69%	68.99%	69.28%	68.12%
cardio-tocography	98.59%	91.22%	92.63%	79.78%
heart (SPECTF)	87.5%	86.25%	85%	82.5%

Hasil pengujian uji-t menunjukkan bahwa metode yang diusulkan dapat meningkatkan kinerja k-NN dalam hal akurasi secara signifikan pada semua dataset yang ditandai dengan nilai p uji-t <0,05. Hasil Uji-t dapat dilihat pada Tabel 4.

Hasil percobaan yang dijelaskan pada Tabel 3 menyatakan bahwa metode yang diusulkan lebih unggul jika dibandingkan dengan algoritma pemilihan fitur lainnya di semua dataset yang dibandingan. Hasil *backward elimination* dan *forward selection* sedikit lebih rendah 0,37% - 5,96% jika dibandingkan dengan algoritma genetika, dan hasil terendah diperoleh dengan seleksi fitur *greedy*.

Berdasarkan hasil percobaan, untuk mengetahui apakah pemilihan fitur dapat meningkatkan performansi dalam klasifikasi dataset medis secara signifikan. t-Test Paired Two Sample for Means digunakan dalam hasil yang diperoleh dari semua algoritma pemilihan fitur.

Hasil pengujian uji-t menunjukkan bahwa pemilihan fitur dapat meningkatkan performansi k-NN dalam hal akurasi secara signifikan pada semua dataset kecuali pemilihan fitur greedy yang ditandai dengan nilai p dari t-Test <0,05. Hasil Uji-t

signifikansi penggunaan fitur seleksi dapat dilihat pada Tabel 5.

Algoritma *k*-nearest neighbors algorithm sangat mudah diimplementasikan (Gorunescu, 2011) dan memiliki akurasi yang tinggi (Harrington, 2012) untuk pengaplikasian yang bervariasi.

Tabel 4. Hasil Pengujian *t-Test* dari Metode yang Diusulkan dibandingkan dengan algoritma k-NN

	Proposed Method	Normal
Mean	88.684	80.494
Variance	125.98713	170.19713
Observations	5	5
Pearson Correlation	0.995007081	
Hypothesized Mean Difference	0	
df	4	
t Stat	8.376046049	
P(T<=t) one-tail	0.000555628	
t Critical one-tail	2.131846786	
P(T<=t) two-tail	0.001111256	
t Critical two-tail	2.776445105	

Tabel 5. Hasil *t-Test* Pengujian Algoritma Genetika dengan Seleksi Fitur Lain

Algoritma	*p* Value of *t*-Test	Results
Genetic Algorithms	0.0011	Sig. (p<0.05)
Forward Selection	0.02	Sig. (p<0.05)
Backward Elimination	0.01	Sig. (p<0.05)
Greedy Feature Selection	0.99	Not Sig. (p>0.05)

Algoritma *k*-NN sangatlah baik karena kelebihan k-NN dianggap sebanding dengan algoritma yang jauh lebih kompleks seperti *neural network* atau *support vector machine*. Dari hasil penelitian ini dapat disimpulkan bahwa k-NN yang dioptimasi parameter yang digabungkan dengan algoritma genetika sebagai fitur seleksi lebih unggul jika dibandingkan dengan algoritma seleksi fitur lainnya pada lima dataset medis yang dijadikan patokan peneliti lain.

4. Kesimpulan

Algoritma genetika diterapkan untuk memilih fitur dan mengoptimalkan parameter *k* untuk tetangga terdekat *k* untuk meningkatkan akurasi dari lima kumpulan data medis yang dijadikan patokan. Metode yang diusulkan terbukti efektif meningkatkan akurasi, dan selanjutnya perbedaan hasil pengujian antar kelima dataset menghasilkan perbedaan yang signifikan.

Perbandingan algoritma seleksi fitur diusulkan untuk membandingkan akurasi hasil antara algoritma genetika, *forward selection*, *backward elimination* dan *greedy feature selection*. Algoritma genetika terbukti memiliki akurasi tertinggi dibandingkan dengan algoritma pemilihan fitur lainnya.

Dalam penelitian ini, secara umum algoritma genetika diterapkan untuk menyeleksi fitur dan mengoptimalkan parameter untuk meningkatkan akurasi lima dataset medis yang sudah dijadikan patokan evaluasi algoritma oleh beberapa penelitian lain. Dalam penelitian selanjutnya, beberapa hal dapat diterapkan untuk menyempurnakan penelitian, yaitu menggunakan algoritma lain untuk optimasi parameter atau metode lain untuk mereduksi dimensionalitas dataset medis.

Referensi

Abe, S. (2005). Modified Backward Feature Selection by Cross Validation. (pp. 163-168). Bruges: European Symposium on Artificial Neural Networks.

Abe, S. (2010). *Support Vector Machine for Pattern Classification* (Second Edition ed.). New York: Springer London.

Amato, F., Lopez, A., Pena-Mendez, E. M., Vanhara, P., Hampi, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine, 11*(2), 47-58.

Antal, B., & Hajdu, A. (2014). An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems, 60*, 20-27.

Ayres-de-campos, D., Bernardes, J., Garrido, A., Marques-de-Sa, J., & Pereira-Leite, L. (2000). SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms. *The Journal of Maternal-Fetal Medicine, 9*, 311-318.

Babu, G. S., & Suresh, S. (2013). Meta-cognitive RBF network and its projection based learning algorithm for classification problems. *Applied Soft Computing Journal, 13*(1), 654-666.

Bharti, K. K., & Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science, 5*(2), 156-169.

Blanchet, F. G., Legendre, P., & Borcard, D. (2008). Forward Selection of Explanatory Variables. *Ecology, 89*(9), 2623-2632.

Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation , 5*(1), 17-26.

Chang, P.-C., Lin, J.-J., & Liu, C.-H. (2012). An attribute weight assignment and particle swarm optimization algorithm for medical database classifications. *Computer Methods and Programs in Biomedicine, 107*(3), 382-392.

Derksen, S., & Keselman, H. J. (1992). Backward, Forward and Stepwise Automated Subset Selection Algorithms. *British Journal of Mathematical and Statistical Psychology, 45*, 265-282.

Dyer, E. L., Sankaranarayanan, A. C., & Baraniuk, R. G. (2013). Greedy Feature Selection for Subspace Clustering. *Journal of Machine Learning Research, 14*, 2487-2517.

Farahat, A. K., Ghodsi, A., & Kamel, M. S. (2013). Efficient Greedy Feature Selection for Unsupervised Learning. *Knowledge Information System, 35*, 285-310.

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques.* Verlag Berlin Heidelberg: Springer.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* San Fransisco: Morgan Kauffman.

Harrington, P. (2012). *Machine Learning in Action.* New York: Manning Publication.

Holland, J. H. (1975). *Adaption in Natural and Artificial Systems.* Cambridge: MIT Press.

Inbarani, H. H., Azar, A. T., & Jothi, G. (2014). Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Computer Methods and Programs in Biomedicine, 113*(1), 175-185.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. *Procedia Technology, 10*, 85-94.

Jain, A., & Zongker, D. (1997). Feature Selection: Evaluation, Application and Small Sample Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(2), 153-158.

Jirapech-Umpai, T., & Aitken, S. (2005). Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics, 6*, 148.

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated Cardiac SPECT Diagnosis. *Artificial Intelligence in Medicine, 23*, 149-169.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc.

Larose, D. T. (2006). *Data Mining Methods and Models.* New Jersey: John Wiley & Sons, Inc.

Liu, Z., Chai, T., & Tang, J. (2015). Multi-frequency signal modeling using empirical mode decomposition and PCA with application to mill load estimation. *Neurocomputing, 169*, 392-402.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook* (Second Edition ed.). New York: Springer.

Man, K. F., Tang, K. S., & Kwong, S. (1996). Genetic Algorithms: Concepts and Applications. *IEEE Transactions on Industrial Electronics, 43*(5), 519-534.

Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research, 43*(4), 570-577.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks, 21*(2), 427-436.

Nowe, A. (2014). *Genetic Algorithms* (Encyclopedia of Astrobiology ed.). Berlin: Springer.

Prasetio, R. T., & Pratiwi. (2015). Penerapan Teknik Bagging pada Algoritma

Klasifikasi untuk Mengatasi Ketidakseimbangan Kelas pada Dataset Medis. *Informatika, 2*(2), 395-403.

Prasetio, R. T., & Riana, D. (2015). A Comparison of Classification Methods in Vertebral Column Disorder with the Application of Genetic Algorithm and Bagging. Bandung: IEEE.

Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., & Jain, A. K. (2000). Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation, 4*(2), 164-171.

Setiyorini, T., & Wahono, R. S. (2015). Penerapan Metode Bagging untuk Mengurangi Data Noise pada Neural Network untuk Estimasi Kuat Tekan Beton. *Journal of Intelligent Systems, 1*(1), 37-42.

Shah, S., & Kusiak, A. (2007). Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine, 37*(2), 251-261.

Shilaskar, S., & Ghatol, A. (2013). Dimensionality Reduction Techniques for Improved Diagnosis of Heart Disease. *International Journal of Computer Applications , 61*(5), 1-8.

Subbulakhsmi, C. V., & Deepa, S. N. (2015). Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier. *The Scientific World Journal, 2015*, 1-12.

Suguna, N., & Thanushkodi, K. (2010). An Improved k-Nearest Neigbor Classification Using Genetic Algorithm. *IJCSI International Journal of Computer Science, 7*(2), 18-44.

Unal, Y., & Kocer, E. (2013). Diagnosis of Pathology on the Vertebral Column with Backpropagation and Naive Bayes Classifier. (pp. 278-281). Turkey: IEEE.

Vafaie, H., & Imam, I. F. (1994). Feature Selection Method: Genetic Algorithms vs Greedy-like Search. Louisville: Proceedings of the 3rd International Fuzzy Systems and Intelligent Control Conference.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Technique* (Third Edition ed.). Amsterdam: Elsevier Inc.

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining.* Boca Raton: Taylor & Francis Group, LLC.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., & Yang, Q. (2008). *Top 10 Algorithms in Data Mining.* London: Springer-Verlag.

Yang, J., & Honavar, V. (1998). Feature Subset Selection Using a Genetic Algorithm. *Feature Extraction, Construction and Selection*, 117-136.