PENERAPAN METODE SELEKSI FITUR UNTUK MENINGKATKAN HASIL DIAGNOSIS KANKER PAYUDARA

Elvira Sukma Wahyuni
Fakultas Teknologi Industri, Program Studi Teknik Elektro
Universitas Islam Indonesia
Email: elvira.wahyuni@mail.uii.ac.id

ABSTRAK

Tujuan utama penelitian ini adalah untuk meningkatkan peforma klasifikasi pada diagnosis kanker payudara dengan menerapkan seleksi fitur pada beberapa algoritme klasifikasi. Penelitian ini menggunakan database kanker payudara *Wisconsin Breast Cancer Database* (WBCD). Metode seleksi fitur F-score dan Rough Set akan dipasangkan dengan beberapa algoritme klasifikasi yaitu SMO (*Sequential Minimal Optimization*), Naive Bayes, Multi layer Perceptron, dan C4.5. Penelitian ini menggunakan 10 *fold cross validation* sebagai metode evaluasi. Hasil penelitian menunjukkan algoritme klasifikasi MLP dan C4.5 mengalami peningkatan peforma klasifikasi secara signifikan setelah dipasangkan dengan seleksi fitur rough set dan F-score, Naive Bayes menunjukan peforma terbaik ketika dipasangkan dengan metode seleksi fitur F-score saja, sedangkan SMO tidak menunjukkan peningkatan peforma klasifikas ketika dipasangkan pada kedua seleksi fitur.

Kata kunci: kanker payudara, seleksi fitur, klasifikasi.

ABSTRACT

The objective of this study is to improve the performance classification of breast cancer diagnosis by applying feature selection on various classification algorithms . This study uses a database of Wisconsin Breast Cancer Database (WBCD). Feature selection methods F -score and Rough Set will be paired with various classification algorithms i.e. SMO (Sequential Minimal Optimization), Naive Bayes , Multi Layer Perceptron , and C4.5. 10-fold cross validation is use as an evaluation method. The results showed MLP and C4.5 has improved performance classification significantly when paired with rough sets and F -score feature selection methods, Naive Bayes showed best Performance when paired with F -score feature selection method, whereas SMO did not show improved performance when paired on both feature selection.

Keywords: *breast cancer, fitur selection, classification.*

1. PENDAHULUAN

Kanker payudara (*Carcinoma mammae*) didefinisikan sebagai suatu penyakit *neoplasma* ganas yang berasal dari *parenchyma*. Penyakit ini oleh World Health Organization (WHO) dimasukkan ke dalam *International Classification of Diseases* (ICD) dengan kode nomor 17 [1]. Frekuensi kasus penyakit ini relatif tinggi di negara maju dan merupakan jenis kanker yang banyak diderita dari jenis kanker lainnya. Di Indonesia, kanker payudara menempati peringkat kedua setelah kanker servik [2]. Menurut data terakhir WHO, angka kematian karena kanker payudara di Indonesia mencapai 20.052 atau sebanyak 1,41% dari seluruh kematian atau angka kematian disesuaikan dengan usia adalah 2.025 per 100.000 penduduk [3]. Kunci untuk bertahan hidup penderita kanker payudara adalah mendeteksi kanker payudara sedini mungkin, sebelum kanker tersebut memiliki kesempatan untuk menyebar [2].

Seiring dengan kemajuan teknologi informasi terutama dalam bidang kecerdasan buatan, teknik *machine learning* diperkenalkan untuk membantu meningkatkan kemampuan pendeteksian otomatis. Dengan bantuan sistem ini, kemungkinan kesalahan diagnosis yang dilakukan oleh para ahli dapat dihindari, dan data medis dapat diperiksa dalam kurun waktu yang singkat serta lebih rinci [4]. Teknik statistik dan teknik kecerdasan buatan telah digunakan untuk memprediksi kanker payudara oleh beberapa peneliti. Tujuan dari teknik ini adalah untuk menetapkan identifikasi pasien ke dalam grup jinak (yang tidak memiliki kanker payudara) atau kelompok ganas (yang terbukti kuat memiliki kanker payudara) [5].

Data medis yang berdimensi tinggi merupakan salah satu kendala dalam penerapan teknik *machine learning* karena akan memberikan efek negatif terhadap proses analisis. Untuk menangani data medis berdimensi tinggi tersebut, mereduksi fitur menjadi hal yang sangat penting. Dengan pengurangan fitur

tidak mengakibatkan kemampuan diskriminatif menjadi memburuk, bahkan sebaliknya terdapat banyak keuntungan diantaranya dapat menghindari *over-fitting*, mengurangi kompleksitas analisis data dan meningkatkan kinerja analisis data [6]. Salah satu usaha untuk mengurangi fitur data yang berdimensi tinggi adalah dengan menggunakan seleksi fitur, seleksi fitur merupakan bagian dari *preprocessing* pada proses klasifikasi. Pemilihan fitur sangat mempengaruhi keakuratan klasifikasi dalam kasus kanker payudara.

Pada penelitian ini dua metode fitur seleksi akan diterapkan dan kemudian akan diuji cobakan pada beberapa algoritme klasifikasi yang berbeda, rough set merupakan seleksi fitur yang dapat mengidentifikasi fitur-fitur yang signifikan dan menghilangkan fitur-fitur yang tidak relevan untuk menghasilkan model pembelajaran yang baik, sehingga dapat mengurangi dimensi data tanpa kekurangan informasi yang terkandung dalam kumpulan data tersebut. F-score sendiri merupakan seleksi fitur dengan teknik sederhana yang mengukur diskriminasi dua set bilangan real, sehingga fitur yang memiliki nilai F-score rendah dianggap memeiliki kemampuan diskriminatif yang rendah pula begitu pula sebaliknya fitur yang memiliki nilai F-score tinggi juga akan memiliki kemampuan diskriminatif yang tinggi pula. Dalam penelitian terdahulu diketahuai seleksi fitur rough set [7] dan seleksi fitur F-score [4] memiliki kemampuan yang sangat baik dalam memilih fitur-fitur yang signifikan terhadap klasifikasi.

1.1 Penelitian Yang Berhubungan

Beberapa penelitian yang sama mengenai seleksi fitur telah dilakuakn diantaranya dalam penelitian [8], penelitian ini mengusulkan sebuah metode seleksi fitur yang diberi nama SVM-FuzCocs. Metode tersebut mengatasi ruang fitur berdimensi tinggi dengan penilaian kualitas fitur berdasarkan keanggotaan fuzzy hasil keluaran dari SVM. Dan hasilnya menunjukkan akurasi klasifikasi dan pengurangan dimensi yang cukup memuaskan. Selain itu, metode ini memiliki kebutuhan komputasi yang cukup rendah.

Penelitian [9] menerapkan *t-test* dan *p-value* untuk meredukasi ruang fitur. Dan hasil penelitian ini menunjukkan bahwa dengan adanya penerapan kedua seleksi fitur tersebut dapat meningkatkan kecepatan proses klasifikasi tanpa menurunkan hasil klasifikasi. Hal ini membuktikan bahwa penggunaan seleksi fitur tidak hanya ditujukan untuk peningkatan peforma klasifikasi, namun juga menurunkan beban komputasi klasifikasi.

Penelitian [10] menerapkan ekstraksi fitur Principal Component Analysis (PCA) dan secara lebih rinci tiga algoritma terbaik dari PCA yaitu *Scree Test*, *Cumulative Varience* dan KG *rule*, digunakan sebagai seleksi fitur dan Artificial Neural Network (ANNs) digunakan sebakai *classifier*-nya. Pada penelitian ini menunjukan rata-rata akurasi klasifikasi terbaik deicapai oleh seleksi fitur *Cumulative Varience* sebesar 95,68%. Hal ini membuktikan bahwa ketiga algoritma seleksi fitur terbaik yang dimiliki PCA mampu meningkatkan akurasi klasifikasi dengan metode ANNs.

Penelitian [11] menerapkan beberapa metode klasifikasi dan metode seleksi fitur diantaranya Support Vector MachineS (SVM), K-nearest neighbours dan probabilistic neural networks *classifiers* akan dikombinasikan dengan signal-to-noise ratio *feature ranking*, dan equential forward selection sebagai fitur seleksi serta principal component analysis *feature extraction*. Hasil penelitian ini menunjukkan pencapaian akurasi antara 98,80% dan 96,33% dengan SVM sebagai *classifier* yang dominan.

1.2 Konsep dasar teori

1.2.1 Seleksi fitur

Konsep dasar metode seleksi fitur yang digunakan pada penelitian ini akan dijelaskan sebagai berikut.

a. F-score

F-score adalah teknik sederhana yang mengukur diskriminasi dua set bilangan real. Dengan training vektor x_k, k = 1,. . . , m, jika jumlah *intance* positif dan negatif n^+ dan n^- masing-masing, maka F-skor dari fitur ke-i didefinisikan pada persamaan (1):

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (1)$$

Dimana masing-masing $\bar{x}$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ adalah rata-rata dari fitur ke-*i* dari keseluruhan positif dan negatif keseluruhan dataset; $x_{k,i}^{(+)}$ adalah fitur ke-*i* dari positif *instance* ke-*k*, dan $x_{k,i}^{(+)}$ adalah fitur ke-i dari negatif *instance* ke-*k* negatif. Diskriminasi antara positif dan negatif set dindikasikan oleh numerator, dan denominator diindikasikan satu di dalam setiap dua set. sebuah fitur yang memiliki nilai F-score yang besar adalah fitur yang sangat dikriminatif. Kemudian, dalam penelitian ini menggunakan F-score untuk kriteria penyeleksian fitur [12].

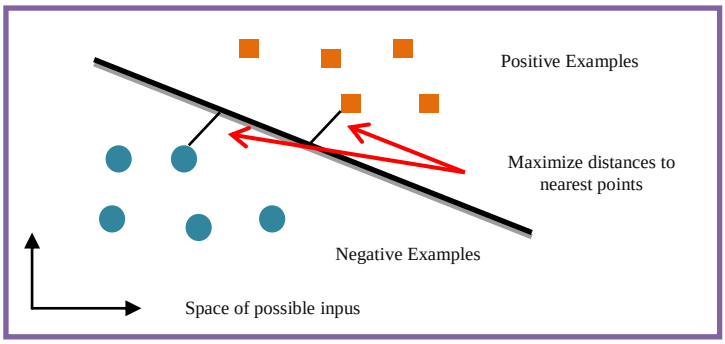
b. Rough Set

Teori *rough set* adalah sebuah *tool* matematika cerdas yang di perkenalkan oleh Prof. Pawlak pada tahun 1982 untuk menangani ketidakpastian dan ketidaklengkapan. Hal tersebut didasarkan pada konsep *upper* dan *lower approximation* dari suatu himpunan, model dan ruang himpunan. keunggulan utama yang dimiliki *rough set* adalah tidak memerlukan informasi awal atau informasi tambahan mengenai data. Salah satu aplikasi utama dari teori *rough set* adalah *atribute reduction*. Reduksi atribut diperoleh dengan membandingkan kesetaraan hubungan yang dibangun oleh himpunan atribut. Dengan menggunakan tingkat ketergantungan sebagai ukuran [7].

1.2.2 Metode Klasifikasi

a. SMO

SMO adalah sebuah algoritme yang mengatasi permasalahan optimisasi *Quadratic Programming* (QP) pada SVM (*support vector machine*). SMO mampu memperkecil permasalahan QP dan dapat memperkecil waktu optimisasi [13]. SVM sendiri adalah metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimizaton* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada input *space* [14], ilustrasi pencarian *hyperplane* terbaik dapat dilihat pada Gambar 1.



Gambar 1. *Linear Support Vector Machine*

b. Naive Bayes

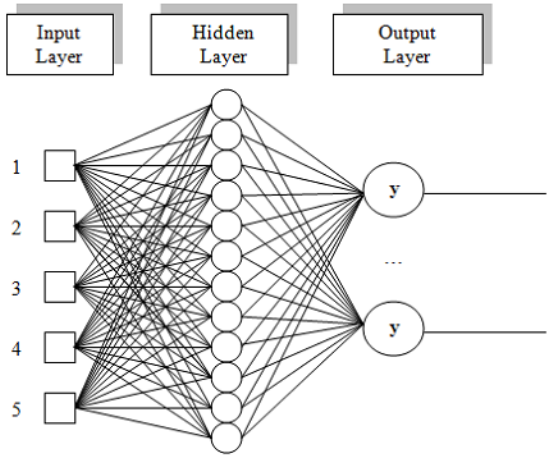
Naive Bayesian adalah metode klasifikasi yang berdasarkan probabilitas, dengan asumsi bahwa setiap variabel X bersifat bebas (*independent*). Dengan kata lain, Naïve Bayesian mengansumsikan bahwa keberadaan sebuah atribut tidak ada kaitannya dengan beradaan atribut yang lain. Jika diketahui X adalah data sampel dengan kelas (label) yang tidak diketahui, H merupakan hipotesa bahwa X adalah data dengan klas (label) C, P(H) adalah peluang dari hipotesa H, P(X) adalah peluang data sampel yang diamati, maka P(X|H) adalah peluang data sampel X, bila diasumsikan bahwa hipotesa H benar (valid).
Karena asumsi atribut tidak saling terkait (*conditionally independent*), maka P(X|Ci) dapat didefinisikan pada persamaan (2):

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (2)$$

Jika P(X|Ci) diketahui maka klas dari data sampel X dapat didekati dengan menghitungg P(X|Ci)*P(Ci). Klas Ci dimana P(X|Ci)*P(Ci) maksimum adalah klas dari sampel X.

c. Multi Layer Preceptron (MLP)

Algoritme MLP merupakan algoritme yang mengadopsi cara kerja jaringan saraf pada mahluk hidup (*artificial neural network*). Algoritme ini dikenal handal karena proses pembelajaran yang mampu dilakukan secara terarah. Pembelajaran yang dilakukan adalah dengan peng-*update*-an bobot balik (*backpropagation*). Penetapan bobot yang optimal akan menghasilkan klasifikasi yang tepat [15]. Arsitektur MLP dapat dilihat pada Gambar 2.



Gambar 2. Contoh Arsitektur MLP [15]

d. C4.5

C4.5 adalah sebuah *decision tree* yang digunakan untuk klasifikasi dengan konsep *information entropy*. Untuk menghasilkan sebuah *pruned tree* C4.5. pembuatan keputusan dilakukan dengan men-*splitting* setiap atribut data kedalam subset yang lebih kecil untuk memeriksa *entropy* yang berbeda, dan memilih atribut dengan *information gain* tertinggi. *Splitting* dihentikan ketika menemukan subset *instance* yang dimasukkan kedalam kelas yang sama, dan dengan demikian *leaf node* akan dibuat. Jika tidak ada *leaf node* yang ditemukan, C4.5 menciptakan simpul tujuan lebih tinggi berdasarkan nilai kelas yang diharapkan [16].

1.2.3 Evaluasi Performa

a. Akurasi, Sensitivitas, Spesifisitas dan ROC curves

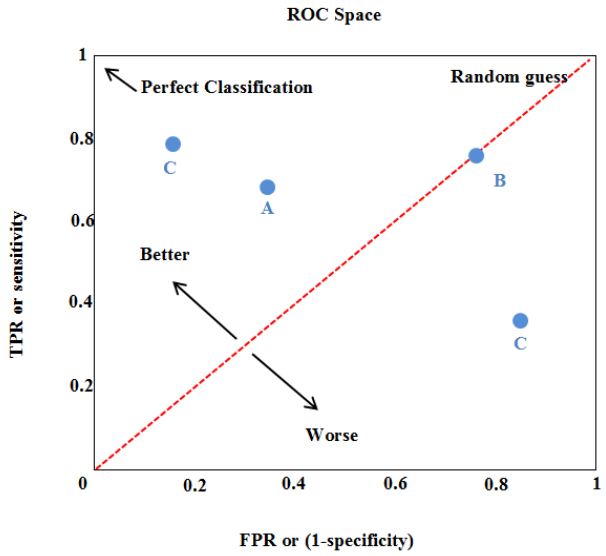
Dalam penelitian ini peforma masing algoritme klasifikasi terhadap dua seleksi fitur akan diukur berdasarkan *accuracy*, *sensitivity*, *specificity* dan ROC *curves*. Dengan formula pada persamaan (3)-(5).

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Sensitivity = \frac{FN}{FN + TN} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (5)$$

Area Under the ROC Curve (AUC) digunakan sebagai metode evaluasi, dimana AUC menghitung luas daerah di bawah kurva ROC. AUC memiliki nilai dengan rentang antara 0,0–1,0, semakin nilai AUC mendekati nilai 1 maka menunjukkan semakin tinggi pula keakuratan klasifikasi. Gambar 3 memperlihatkan contoh kurva ROC.



Gambar 3. Kurva ROC

b. t-test

Untuk membandingkan performa masing-masing metode seleksi fitur F-score dan *rough set*, maka akan dilakukan uji *t-est* untuk melihat taraf signifikan perbedaan performa yang dihasilkan. Pada penelitian ini akan digunakan *paired sample t-test* untuk mengujikan *sample* berpasangan, yaitu data yang sama namun mendapatkan perlakukan yang berbeda. Pengujian *paired sample t-test* menggunakan formula (6).

$$t = \frac{m_d}{\sqrt{\sigma_d^2 / k}} \quad (6)$$

2. METODOLOGI PENELITIAN

2.1 Data Set

Data yang digunakan pada penelitian ini adalah *wisconsin breast cancer database* (WBCD) yang diambil dari UCI *machine learning repository* yang bersumber dari University of Wisconsin Hospitals, Madison dari Dr. William H. Wolberg (http://archive.ics.uci.edu/ml/datasets.html).

Dataset ini umum digunakan oleh para peneliti yang menggunakan *Machine learning* sebagai metode klasifikasi kanker payudara, Dataset berisi 699 sampel yang diambil dari *needle aspirates* dari jaringan kanker payudara manusia, dimana terdapat 16 *instance* yang memiliki *missing value*. karena *missing value* yang ditemukan dalam jumlah yang sangat kecil dibandingkan jumlah keseluruhan data maka 16 *instance* tersebut dibuang sehingga jumlah *instance* yang digunakan sebanyak 683.

Terdiri dari sembilan fitur, yang masing-masing direpresentasikan sebagai integer antara 1-10 dapat dilihat pada Tabel 1.

Tabel 1. Fitur WBCD

Label	*Atribut*	*Domain*
C1	Clump Thickness	1-10
C2	Uniformity of Cell Size	1-10
C3	Uniformity of Cell Shape	1-10
C4	Marginal Adhesion	1-10
C5	Single Epithelial Cell Size	1-10
C6	Bare Nuclei	1-10
C7	Bland Chromatin	1-10
C8	Normal Nucleoli	1-10
C9	Mitoses	1-10

2.2 Alur Penelitian

a. Seleksi Fitur

Metode reduksi oleh *rough set* diterapkan pada *full* fitur *dataset* WBCD. *Genetic algorithm* dipilih sebagai algoritme pencarian set reduksi. Reduksi *rough set* menghasilkan subset-subset kombinasi fitur terbaik berdasarkan *discernibility*, dapat dilihat pada Tabel 2. Selanjutnya subset terpilih diperkecil dengan memilih subset optimal dengan menggunakan strategi "*combination filtering*" [11]. "*combination filtering*" merupakan teknik pemilihan subset optimal berdasarkan subset yang mengandung atribut *strong* dan *weak relevancy* dengan cara menghitung korelasi antara atribut kondisi dengan atribut tujuan, nilai korelasi masing-masing atribut terhadap kelas tujuan diperlihatkan pada Tabel 3 . Hal tersebut dilakukan atas dasar bahwa tidak hanya atribut yang memiliki *strong relevancy* yang dapat membentuk subset optimal terkadang atribut dengan *weak relevancy* juga dapat meningkatkan akurasi [17]. Hasil reduksi seleksi fitur Rough set yang terpilih dapat dilihat pada Tabel 4.

Seleksi fitur menggunakan F-score dilakukan dengan menghitung nilai F-score masing-masing fitur menggunakan Persamaan (1), nilai F-score masing-masing atribut diperlihatkan pada Tabel 5. Kemudian fitur dalam tiap subset diurut secara menurun berdasarkan *ranking* nilai F-score. Selanjutnya akan dibentuk subset baru dengan menggabungkan beberapa kemungkinan kombinasi fitur berdasarkan nilai F-score terbaik. Langkah pembentukan subset baru berdasarkan nilai F-score akan dijelaskan sebagai berikut.

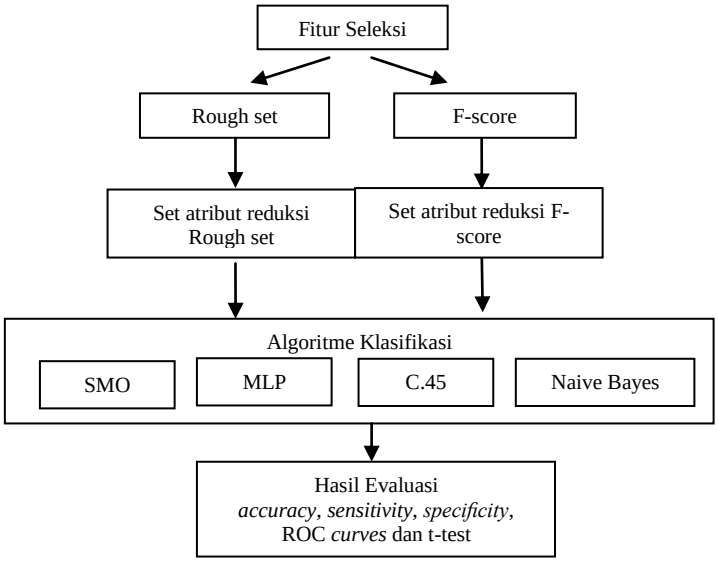
Misalkan C merupakan fitur yang terdapat pada subset D_1, dimana C_i merupakan indeks fitur ke-*i*. M_i adalah nilai F-score C_i. N adalah *ranking* C_i berdasarkan M_i dimana N = 1 . . . . .n, n merupakan jumlah total fitur. Maka pengurutan fitur di dalam setiap subset adalah $D_1 = \{M_1. . . . .M_n\}$. subset baru yang dapat dibentuk pertama kali adalah kombinasi dua urutan M_n tertinggi yaitu $D_2 = \{M_1, M_2\}$, subset berikutnya adalah N = N+1 hanya jika N < n. Misalkan n = 5, maka subset yang terbentuk adalah $D_2 = \{M_1, M_2\}$, $\{M_1, M_2, M_3\}$, $\{M_1, M_2, M_3, M_4\}$. Set atribut hasil seleksi fitur F-score diperlihatkan pada Tabel 6.

b. Setting Parameter

Beberapa algoritme klasifikasi menghendaki pengaturan pada parameter tertentu. Algoritme SMO yang diterapkan pada penelitian ini menggunakan RBF (*Radial Basis Function*) kernel ada dua parameter yang harus ditentukan yaitu C dan γ . Untuk mencari parameter C dan γ yang optimum penelitian ini menerapkan teknik *grid search* dengan 10 fold cross validation dengan *grid space* $\log_2$ C {1,2,3. . . . , 16} dan $\log_2$ γ {-5,-4,. . . . , 2}. Algoritme C.45 menggunakan standar *cofidence factor* (25%). MLP menggunkan tiga leyer, yang terdiri dari *input layer* (28 neuron), satu *hidden layer* (15 neuron), dan satu *output layer* (dua neuron). Penyesuaian bobot dilakukan pada 500 siklus.

c. Klasifikasi

Pada tahap ini subset fitur hasil seleksi akan diklasifikasi dengan menggunakan beberapa algoritme yaitu SMO, MLP, C4.5 dan Naive Bayes, secara garis besar skema alur penelitian dapat dilihat pada Gambar 4. Klasifikasi SVM dilakukan dengan bantuan perangkat lunak Weka.



Gambar 4. Skema alur penelitian

Tabel 2. Hasil set atribut yang teridentifikasi oleh rough set

No Set Atribut	*Set Atribut*
1	{C1, C2, C5, C6}
2	{C1, C3, C6, C7}
3	{C1, C4, C6, C8}
4	{C1, C5, C6, C8}
5	{C1, C2, C6, C8}
6	{C3, C5, C6, C8}
7	{C1, C4, C6, C7}
8	{C1, C3, C6, C8}
9	{C2, C3, C4, C6, C7}
10	{C3, C4, C6, C7, C9}
11	{C1, C3, C4, C6, C9}
12	{C1, C2, C3, C4, C6}
13	{C2, C5, C6, C7, C9}
14	{C1, C2, C5, C6, C9}
15	{C1, C2, C4, C6, C9}
16	{C2, C5, C6, C7, C8}
17	{C2, C5, C6, C8, C9}
18	{C2, C4, C5, C6,C7}
19	{C2, C4, C5, C6, C8}
20	{C5, C6, C4, C8, C9}

Tabel 3. Nilai korelasi atribut

No	*Atribut*	*Nilai Korelasi*
1	C1	0,712
2	C2	0,820
3	C3	0,821
4	C4	0,706
5	C5	0,690
6	**C6**	**0,822**
7	C7	0,489
8	C8	0,718
9	**C9**	**0,423**

Tabel 4. Subset fitur yang terpilih

No Subset	*Subset Fitur*	*Jumlah Fitur*
1	{C2, C5, C6, C7, C9}	5
2	{C2, C5, C6, C8, C9}	5
3	{C5, C4, C6, C7, C9}	5
4	{C1, C2, C5, C6, C9}	5
5	{C1, C3, C4, C6, C9}	5
6	{C1, C2, C4, C6, C9}	5
7	{C5, C6, C7, C8, C9}	5

Tabel 5. Nilai F-score masing-masing fitur

Label	*Nilai F-score*	*Peringkat*
C1	1,112691644	5
C2	1,857298354	3
C3	1,920505411	2
C4	0,885539239	7
C5	0,837800748	8
C6	1,936842827	1
C7	1,302362589	4
C8	0,949633087	6
C9	0,18839	9

Tabel 6. Sembilan set atribut yang disusun berdasarkan nilai F-score

No Set Atribut	*Nilai F-Score*
1	C6
2	C6, C3
3	C6,C3,C2
4	C6,C3,C2,C7
5	C6,C3,C2,C7,C1
6	C6,C3,C2,C7,C1,C8
7	C6,C3,C2,C7,C1,C8,C4
8	C6,C3,C2,C7,C1,C8,C4,C5
9	C6,C3,C2,C7,C1,C8,C4,C5,C9

3. HASIL DAN PEMBAHASAN

3.1 Eksperimen dengan menggunakan fitur seleksi Rough set

Tabel 7 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme SMO dan seleksi fitur Rough set. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #2 yaitu {C2, C5, C6, C8, C9}.

Tabel 7. Seleksi fitur Rough set dan algoritme klasifikasi SMO

No Set	*SMO*			
Atribut	*Akurasi*	*Sensitivitas*	*Spesifisitas*	*ROC AUC*
#1	96,7789	0,983945	0,939271	0,968
#2	**96,7789**	**0,986175**	**0,935743**	**0,969**
#3	96, 4861	0,979452	0,938776	0,964
#4	96, 6325	0,983908	0,935484	0,967
#5	96,6325	0,979499	0,942623	0,965
#6	96,1933	0,970721	0,945607	0,958
#7	96,4061	0,977273	0,942387	0,963

Tabel 8 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme MLP dan seleksi fitur Rough set. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #5 yaitu {C1, C3, C4, C6, C9}.

Tabel 8. Seleksi fitur Rough set dan algoritme klasifikasi MLP

No Set	*MLP*			
Atribut	*Akurasi*	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	95,004	0,970588	0,937759	0,986
#2	95,754	0,962138	0,948718	0,987
#3	95, 4612	0,96614	0,933333	0,985
#4	95,002	0,961712	0,92887	0,986
#5	**96,0464**	**0,970655**	**0,941667**	**0,984**
#6	95,9004	0,970588	0,937759	0,981
#7	94,8755	0,961625	0,925	0,984

Tabel 9 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme C4.5 dan seleksi fitur Rough set. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #4 yaitu {C1, C2, C5, C6, C9}.

Tabel 9. Seleksi fitur Rough set dan algoritme klasifikasi C4.5

No Set Atribut	*C.45*			
	Akurasi	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	92,6794	0,926407	0,927602	0,967
#2	93,265	0,938326	0,921397	0,967
#3	91,9473	0,941043	0,880165	0,962
#4	**93,5578**	**0,94843**	**0,911392**	**0,96**
#5	91,2152	0,922907	0,89083	0,944
#6	93,4114	0,952381	0,900826	0,958
#7	92,3865	0,9375	0,897872	0,905

Tabel 10 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme Naive bayes dan seleksi fitur Rough set. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #2 yaitu {C2, C5, C6, C8, C9}.

Tabel 10. Seleksi fitur Rough set dan algoritme klasifikasi Naive Bayes

No Set Atribut	*Naive Bayes*			
	Akurasi	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	96,6325	0,981693	0,939024	0,992
#2	**96,9235**	**0,986207**	**0,939516**	**0,99**
#3	96,7789	0,981735	0,942857	0,993
#4	96,6325	0,983908	0,935484	0,992
#5	96,6325	0,981693	0,939024	0,993
#6	96,6325	0,986143	0,932	0,994
#7	96,1937	0,975	0,938272	0,99

3.2 Eksperimen dengan menggunakan fitur seleksi F-score

Tabel 11 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme SMO dan seleksi fitur F-score. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #9 yaitu {C1,C2,C3,C4,C5, C6,C7, C8, C9}.

Tabel 11. Seleksi fitur Rough set dan algoritme klasifikasi SMO

No Set Atribut	*SMO*			
	Akurasi	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	90,0439	0,921525	0,860759	0,89
#2	94,2899	0,961276	0,909836	0,94
#3	96,3397	0,983834	0,928	0,965
#4	96,6325	0,990783	0,943775	0,967
#5	97,3646	0,984091	0,954733	0,973
#6	97,2182	0,984055	0,95082	0,972
#7	97,3646	0,984091	0,954733	0,973
#8	97,0717	0,984018	0,946939	0,971
#9	**97,6574**	**0,993088**	**0,947791**	**0,979**

Tabel 12 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme MLP dan seleksi fitur F-score. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #8 yaitu {C1,C2,C3,C4,C5, C6,C7, C8}.

Tabel 12. Seleksi fitur Rough set dan algoritme klasifikasi MLP

No Set Atribut	*MLP*			
	Akurasi	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	90,6290	0,918502	0,882096	0,983
#2	94,5827	0,955257	0,927966	0,979
#3	94,5827	0,968326	0,93361	0,986
#4	95,6076	0,961798	0,932773	0,986
#5	95,1684	0,970455	0,930041	0,988
#6	95,3148	0,959821	0,940426	0,991
#7	95,6076	0,970455	0,930041	0,987
#8	**96,1933**	**0,975**	**0,938272**	**0,991**
#9	95,9004	0,961712	0,930612	0,989

Tabel 13 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme C4.5 dan seleksi fitur F-score. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #3 dan #4 yaitu {C1,C2,C3}dan {C1,C2,C3,C4}.

Tabel 13. Seleksi fitur Rough set dan algoritme klasifikasi C4.5

No Set Atribut	*C4.5*			
	Akurasi	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	89,8975	0,915743	0,866379	0,934
#2	94,8755	0,972286	0,908	0,972
#3	**95,1684**	**0,961798**	**0,932773**	**0,966**
#4	**95,1684**	**0,961798**	**0,932773**	**0,966**
#5	93,7042	0,948546	0,915254	0,968
#6	93,5570	0,94843	0,911392	0,966
#7	93,5578	0,94843	0,911392	0,966
#8	93,4114	0,948315	0,907563	0,956
#9	93,4114	0,948315	0,907563	0,956

Tabel 14 memperlihatkan perbandingan hasil akurasi, sensitivitas, spesifisitas dan ROC AUC klasifikasi algoritme Naive bayes dan seleksi fitur F-score. Akurasi, sensitivitas, spesifisitas dan ROC AUC tertinggi diperoleh oleh atribut set nomor #6 {C1,C2,C3,C4,C5,C6}.

Tabel 14. Seleksi fitur Rough set dan algoritme klasifikasi Naive bayes

No Set Atribut	*Naive Bayes*			
	Akurasi	*Sensitiviti*	*Spesifisiti*	*ROC AUC*
#1	89,8975	0,915743	0,866379	0,937
#2	95,6076	0,962054	0,944681	0,988
#3	96,3397	0,979405	0,934959	0,99
#4	96,7789	0,981735	0,942857	0,992
#5	97,3646	0,990783	0,943775	0,994
#6	**97,6574**	**0,990805**	**0,947581**	**0,994**
#7	97,511	0,990805	0,947581	0,994
#8	97,6514	0,993088	0,947791	0,993
#9	97,3646	0,988532	0,947368	0,994

3.3 Perbandingan performa metode klasifikasi

Pada Tabel 15 menunjukkan bahwa pada algoritme klasifikasi MLP dan C4.5 setelah diterapkan metode seleksi fitur baik rough set maupun F-score terjadi peningkatan akurasi, pada algoritme klasifikasi Naive Bayes peninggkatan akurasi hanya pada metode seleksi fitur F-score, sedangkan pada algoritme SMO tidak terjadi peninggkatan akurasi, hasil akurasi tertinggi yang di peroleh sama.

Tabel 15. Perbandingan performa metode klasifikasi sebelum dan sesudah dilakukan seleksi fitur

Classifier	*Akurasi tertinggi*	*Jumlah Atribut*
SMO	97,6574	9
F-Score + SMO	97,6574	9
Rough set +SMO	96,7789	5
MLP	95,9004	9
F-Score + MLP	96,1933	8
Rough set +MLP	96,0464	5
C4.5	93,4114	9
F-Score + C4.5	95,1684	4
Rough set + C4.5	93,5578	5
Naive Bayes	97,3646	9
F-Score + Naive Bayes	97,6574	6
Rough set + Naive bayes	96,9235	5

3.4 Hasil t-test

Hasil perhitungan t-tes untuk memperlihatkan kenaikan hasil diagnosis dengan penerapan seleksi fitur diperlihatkan pada Tabel 16. Menggunakan 95% *confidence level* ($\alpha = 0.05$). pengujian t-test dilakukan hanya pada hasil kalsifikasi yang mengalami peningkatan.

Tabel 16. Hasil t-test

Classifier	*Perbandingan akurasi*	*t-test*	*Keterangan*
SMO	SMO+RoughSet Vs SMO	-	-
	SMO+F-score Vs SMO	-	-
MLP	MLP+RoughSet Vs MLP	9.170261	Signifikan
	MLP+F-score Vs MLP	9.673359	Signifikan
C4.5	C4.5+RoughSet Vs C4.5	2.180661	Signifikan
	C4.5+F-score Vs C4.5	9.479379	Signifikan
Naive Bayes	Naive Bayes+RoughSet Vs Naive Bayes	-	-
	Naive Bayes+F-score Vs Naive Bayes	2,197269	Signifikan

4. KESIMPULAN

Penelitian ini mencoba menerapkan dua seleksi fitur masing-masing Rough set dan F-score dengan beberapa algoritme klasifikasi yaitu SMO, MLP, C4.5, dan Naive Bayes. Hasil penelitian menunjukkan masing-masing algoritme klasifikasi memiliki peforma yang berbeda terhadap masing-masing metode seleksi fitur, dimana MLP dan C4.5 mengalami peninggkatan peforma klasifikasi secara signifikan setelah diterapkan seleksi fitur, Naive Bayes belum menunjukkan peningkatan hasil klasifikasi ketika diterapkan dengan metode seleksi fitur Rough set, sedangkan jika dipasangkan dengan metode seleksi fitur F-score terjadi peningkatan hasil klasifikasi secara signifikan. Algoritme klasifikasi SMO belum menunjukkan adanya peningkatan hasil klasifikasi ketika diterapkan dengan kedua metode seleksi fitur. Dari penelitian yang dilakukan diketahui bahwa metode seleksi dapat meningkatkan hasil diagnosis klasifikasi kanker payudara secara signifikan dengan jumlah fitur yang lebih kecil.

DAFTAR PUSTAKA

[1] "*Breast Cancer"*. Available: http://www.tempo.co.id/medika/arsip/082002/pus-3.htm, Last access 28 Mei 2013.

[2] "*Gejala Kanker Payudara"*. Available: http://www.deherba.com/gejala-gejala-kanker-payudara.html, Last access 28 Mei 2013.

[3] "Deteksi dini kanker Payudara". Available: http://www.daherba.com, Last access 28 Mei 2013.

[4] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, pp. 3240-3247, 2009.

[5] D. Soria, J. M. Garibaldi, E. Biganzoli, and I. O. Ellis, "A Comparison of Three Different Methods for Classification of Breast Cancer Data," in *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on*, 2008, pp. 619-624.

[6] G. Donghai, Y. Weiwei, J. Zilong, and L. Sungyoung, "Undiagnosed samples aided rough set feature selection for medical data," in *Parallel Distributed and Grid Computing (PDGC), 2012 2nd IEEE International Conference on*, 2012, pp. 639-644.

[7] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, pp. 9014-9022, 2011.

[8] S. P. Moustakidis and J. B. Theocharis, "SVM-FuzCoC: A novel SVM-based feature selection method using a fuzzy complementary criterion," *Pattern Recognition*, vol. 43, pp. 3712-3729, 2010.

[9] D. Aijuan and W. Baoying, "Feature selection and analysis on mammogram classification," in *Communications, Computers and Signal Processing, 2009. PacRim 2009. IEEE Pacific Rim Conference on*, 2009, pp. 731-735.

[10] H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on Principal Component Analysis," in *Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium on*, 2010, pp. 1-4.

[11] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, 2010, pp. 114-120.

[12] Y. W. Chen and C. J. Lin, "combining SVMs with Various Feature Selection Strategies."

[13] A. S. N. Dwi Handoko and Arief Budi Witarto, "Support Vector Machine : teori dan aplikasinya dalam bioinformatika."

[14] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika," *Kuliah Umum IlmuKomputer.Com* 2003.

[15] A. Muliantara and I. M. Widiartha, "Penerapan multi layer preceptron dalam anotasi image secara otomatis."

[16] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis," in *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*, 2012, pp. 180-185.

[17] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," in *Machine Learning: Proceedings Of The Eleventh International*, 1994, pp. 121–129.