K-MEANS SEBAGAI EKSTRAKTOR CIRI PADA KLASIFIKASI DATA DENGAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)

Nurul Chamidah
Fakultas Ilmu Komputer, Program Studi Informatika
Universitas Pembangunan Nasional Veteran Jakarta
Email: nurul.chamidah@upnvj.ac.id

ABSTRAK

Besarnya dimensi pada ciri merupakan masalah pada komputasi untuk mengklasifikasi data sehingga diperlukan suatu proses ekstraksi ciri agar dimensinya berkurang dengan cara mengambil hanya informasi yang penting dari ciri. Penelitian ini menggunakan algoritma K-Means untuk mengekstraksi ciri dengan menemukan pola tersembunyi dari setiap kelas kemudian direkonstruksi dengan *fuzzy membership function* dan mendapatkan pola baru. Pola baru yang terbentuk digunakan sebagai ciri abstrak dan dibagi kedalam data latih dan data uji. Pelatihan dilakukan dengan memanfaatkan algoritma Support Vector Machine (SVM) untuk mendapatkan model klasifikasi. Model klasifikasi SVM yang diperoleh kemudian di uji dengan menggunakan data uji untuk memperoleh performa klasifikasi berupa akurasi dan waktu komputasi. Dengan *5-fold cross validation*, metode ini memberikan akurasi yang baik pada *Dataset Liver, Breast Cancer* dan *Heart Disease* yang diperoleh dari *UCI Machine Learning Repository*. Penelitian ini menunjukkan kemampuan K-Means untuk mengekstraksi ciri dari *dataset*. Hasil penelitian ini menujukkan bahwa K-Means sebagai ekstraktor ciri dapat mengurangi waktu komputasi.

Kata kunci: ekstraksi; ciri; klasifikasi; k-means; SVM, akurasi.

ABSTRACT

The high dimension of feature data is one of the problem for computation to classify data, it causing the need for feature extraction to reduce its dimension by retrieving only important information form the features.This research using K-Means Algorithm to extract features and finding hidden pattern form each class then reconstructed by fuzzy membership function and get new pattern. New pattern that obtained is used as abstract features and divided into training data and testing data. Training is done by Support Vector Machine (SVM) algorithm to obtain classification model. SVM classification model is tested using testing data to get classification performance as accuracy and computational time. Based on 5-fold cross validation, this method has a good accuracy using Liver Dataset, Breast Cancer Dataset, and Heart Disease Dataset from UCI Machine Learning Repository. This research shows the ability and capability of K-Means for extracting features in dataset. The result of this research shows that using K-Means as feature extractor can reduce computational time.

Keywords: *extraction; feature; classification; k-means; SVM; accuracy.*

1. PENDAHULUAN

Perkembangan teknologi yang semakin canggih memudahkan manusia mengambil berbagai ciri yang diinginkan pada berbagai masalah dan kasus, menyebabkan pertumbuhan data yang sangat pesat diberbagai bidang. Peningkatan dimensi ciri pada data menyebabkan berbagai analisis data dan klasifikasi menjadi lebih sulit dilakukan karena semakin besar dimensi semakin besar waktu komputasi dan memory yang diperlukan untuk analisis [1].

Fenomena Hughes menyebutkan bahwa peningkatan dimensi ciri tidak menjamin akurasi dari pengenalan [2]. Dari fenomena ini, penelitian-penelitian telah dilakukan untuk mereduksi jumlah ciri seperti dengan pemilihan ciri (*feature subset selection*) dengan perankingan pada *paralinguistic analysis* [3], pada penelitian citra kerusakan mentimum menggunakan *Max-Relevance Min-Redundancy* (MRMR), *Mutual Information Feature Selection* (MIFS), dan *Sequential Forward Selection* (SFS) [4]. Penelitian lainnya mereduksi jumlah ciri dengan melakukan pengurangan dimensi ciri (*dimensionality reduction*) dengan *neural fuzzy* pada data medis [5], *Principal Component Analysis* (PCA) pada data email dan obat[1].

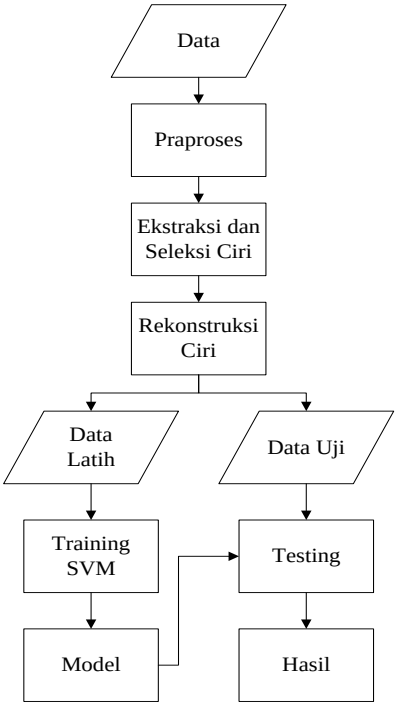
Selain dengan mereduksi jumlah ciri, beberapa peneliti menerapkan ekstraksi dan seleksi ciri sekaligus [6], [7], yakni ciri yang diekstrak dan dipilih merupakan ciri yang baik sehingga akurasi tidak menurun, dan karena telah diekstrak dan dipilih, maka performa klasifikasinya juga menjadi cepat. Ekstraksi dan seleksi ciri dengan K-Means sebagai telah dilakukan [8] untuk mengekstrak data *breast cancer* dan mengklasifikasikan ke dalam dua

kelas dengan SVM dari 32 ciri menjadi 6 ciri. Penelitian Chamidah dan Wasito [9] menerapkan klasifikasi decision tree dan SVM pada tiga kelas dengan data *Cardiotocography* dari 21 ciri menjadi 5. Hasil penelitian menunjukkan signifikansi hasil klasifikasi dan peningkatan performa.

Wolpert dan Mcready [10] menyebutkan, bahwa tidak ada model terbaik secara umum. Setiap kasus memiliki data yang unik dan memiliki penyelesaian yang unik. Oleh karena itu, pada penelitian ini akan diujicobakan klasifikasi dengan K-Means sebagai ekstraktor ciri dan SVM sebagai algoritma klasifikasinya menggunakan *dataset* medis *Liver*, *Breast Cancer*, dan *Heart Disease* untuk mengevaluasi performa ekstraksi dan seleksi ciri pada klasifikasi SVM.

2. METODOLOGI PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1. Penelitian dilakukan dengan (1) mengumpulkan data dan dilakukan suatu (2) praproses. Data yang telah dipraproses kemudian (3) diekstraksi dan seleksi ciri. Hasil ekstraksi dan seleksi kemudian (4) direkonstruksi untuk membentuk ciri abstrak yang merupakan ringkasan informasi yang penting dari ciri. Proses berikutnya, data yang berisi hasil rekonstruksi ciri (4) dibagi ke dalam data latih dan data uji. (5) Data latih digunakan untuk membangun model dengan menggunakan SVM dan (6) data uji untuk memvalidasi model SVM hasil latihan.



Gambar 1. Metode Penelitian

2.1 Data

Data diperoleh dari *UCI Machine Learning Repository* [11]. Data tersebut adalah data *liver* sebanyak 583 *record* yang terdiri dari klasifikasi *liver* sebanyak 167 record, dan *nonliver* sebanyak 416 *record*. Data *breast cancer* dengan total 569 *record* yang terdiri dari 357 *record* berupa kanker jinak dan 212 *record* kanker ganas. Data *heart disease* dengan total 303 *record* yang terdiri dari 160 data sehat dan 143 data sakit. Jumlah ciri dari masing-masing data yang digunakan dalam penelitian ini, untuk data *liver* sebanyak sepuluh ciri, data *breast cancer* sebanyak tiga puluh ciri, sedangkan data *heart disease* sebanyak tiga belas ciri.

2.2 Praproses

Praproses dilakukan dengan melakukan transformasi *min-max* ke dalam *range* 0-1 [12]. Dimana nilai baru *nbaru* diperoleh dengan membandingkan nilai lama *nlama* dikurangi nilai minimum *nmi* pada setiap ciri dengan nilai maksimum *nmaks* pada suatu ciri dikurangi dengan nilai minimumnya *nmi* kemudian dikalikan dengan nilai maksimum baru *maksbaru* dikurangi minimum baru *mibaru* yang diinginkan dan dijumlahkan dengan nilai minimum baru yang diinginkan. Berikut formula normalisasi *min-max*,

$$nbaru = \frac{nlama - nmi}{nmaks - nmi}(maksbaru - mibaru) + mibaru \quad (1)$$

2.3 Ekstraksi dan Seleksi Ciri

Ekstraksi dan seleksi ciri ini dilakukan untuk mendapatkan pola tersembunyi dari tiap kelas secara terpisah dengan Algoritma K-Means. K-Means mengelompokkan ciri-ciri yang paling berpengaruh terhadap suatu kelas dengan mencari jarak terdekat antara pusat (*centroid*) dengan anggota *cluster*-nya.

$$\min_{\bar{x}_1, \bar{x}_2, \ldots \bar{x}_{nK}} \sum_{nk=1}^{nK} \sum_{i \in H_{nk}} \left\| X^i - \bar{x}_{nk} \right\|^2 \quad (2)$$

Dimana nK adalah indeks *cluster*, H_{nk} merupakan *cluster set* ke- nK, $\bar{x}_{nk}$ berupa pusat *cluster* atau *centroid point* dalam *cluster* H_{nk} .

K-Means secara iteratif menghitung jarak terdekat *euclidean distance* dengan mengadaptasi lokasi titik *centroid*. Jumlah *cluster* ditentukan dengan menghitung kriteria Calinski-Harabasz [13], [14]. Indeks Calinski-Harabasz didefinisikan sebagai [15]:

$$CH_C = \frac{SS_B}{SS_W} \times \frac{(N - nK)}{(nK - 1)} \quad (3)$$

$$SS_B = \sum_{x_{nk} \in C} |x_{nk}| \, \|\bar{x}_{nk} - \bar{X}\|^2 \quad (4)$$

$$SS_W = \sum_{x_{nk} \in C} \sum_{x_i \in x_{nk}} \|x_i - \bar{x}_{nk}\|^2 \quad (5)$$

Dimana SS_W adalah jarak antara titik di dalam *cluster* ke *centroid*-nya (*within cluster distance*), SS_B jarak antara *centroid* ke *global centroid* (*between cluster distance*), $|x_{nk}|$ jumlah anggota *cluster* ke- nk, x_{nk} *cluster* ke- nk, $\bar{x}_{nk}$ *centroid cluster* ke- nk, $\bar{X}$ rata-rata *global* dari *dataset*, $C = \{x_1, x_2, .., x_{nK},\}$ yakni *clustering* data X sebanyak N objek ke dalam nK grup dan jumlah grup > 1.

Optimum cluster diperoleh dari nilai CH *index* yang maksimum. suatu hasil *clustering* dikatakan baik jika memiliki *within cluster distance* yang kecil dan *between cluster distance* yang besar Pencarian jumlah *cluster* yang optimal dari setiap *dataset* pada penelitian ini diuji pada jumlah *cluster* pada *range cluster* nK ($2 \leq nK \leq 10$) untuk masing-masing kelas pada setiap *dataset*.

2.4 Rekonstruksi Ciri

Rekonstruksi filakukan setelah mendapatkan jumlah *cluster* optimum dan dilakukan *clustering* dengan K-Means sejumlah *optimum cluster* ini. Pola dari suatu kelas direpresentasikan oleh *cluster-cluster* yang membentuk kelas tersebut sebagai ciri simbolik yang disimbolkan dengan *centroid cluster* pada *cluster* tersebut. Ciri simbolik pada setiap kelas di *dataset* digunakan untuk menghitung similaritas antara data dengan ciri simboliknya. Proses ini berguna untuk mengetahui kecocokan data dengan pola baru yang telah ditemukan. Untuk menghitung similaritas ini digunakan *fuzzy membership function* [8] dari setiap data terhadap pola yang telah diidentifikasi (ciri pada *centroid*). Formulasi *fuzzy membership function* dapat dilihat pada rumus (6). Hasil dari *fuzzy membership function* kemudian digunakan untuk membentuk pola atau ciri simbolik yang dapat dilihat pada formula (7), yakni menghitung semua anggota kelas pada data terhadap fuzzy membership function, sehingga akan terbentuk ciri yang merupakan jumlah *cluster* dari seluruh kelas.

$$fuzzy_{np}(X_j^i) = \begin{cases} 1 - \frac{\left| X_j^{\bar{x}_{np}} - X_j^i \right|}{\max \left| X_j^{\bar{x}_{np}} - X_j^n \right|} & \text{if } (\min(X_j^n) \leq X_j^i \leq \max(X_j^n), \forall n \in H_{np} \\ 0, \; otherwise; \end{cases} \quad (6)$$

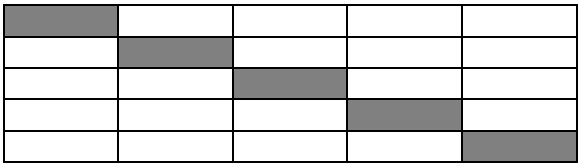
$$Pat_{np} = \frac{1}{D} \sum_{j=1}^{D} fuzzy_{np}(X_j^i), 1 \leq np \leq nK^1 + nK^2 \quad (7)$$

Dimana np :indeks dari pola baru, X_j^i : ciri ke-j dari input asli i, $X_j^{\bar{x}_{np}}$: ciri ke j dari *centroid* $\bar{x}_{np}$ untuk *cluster* H_{np}, nK^1: jumlah pola dari kelas 1, nK^2: jumlah pola dari kelas 2. Pola yang diperoleh dari K-Means merupakan suatu ciri abstrak yang berbeda dari ciri sebelum di lakukan K-Means [8]. Dimensi ciri telah direduksi dengan ciri abstrak baru yang berisi kombinasi dari ciri-ciri sebelumnya dengan informasi ringkas. Kemudian ciri-ciri baru ini digunakan dalam pelatihan dengan SVM untuk memperoleh model klasifikasi yang selanjutnya akan digunakan untuk mengklasifikasi.

2.5 Data Latih dan Data Uji

Data yang telah direkonstruksi dipisah menjadi data latih dan data uji. Data latih digunakan untuk membangun model klasifikasi SVM, dan data uji digunakan untuk menguji model yang telah dibangun. Pembagian data latih dan uji menggunakan *5-fold cross validation*, yakni membagi data ke dalam lima bagian yang sama rata, kemudian secara bergantian mengambil empat bagian sebagai data latih dan satu bagian sebagai data uji. Pembagian data ini diilustrasikan pada

Gambar **2** dimana warna hitam menunjukkan data uji dan warna putih menunjukkan data yang digunakan untuk latihan.



Gambar 2. ***5-Fold Cross Validation***

2.6 Klasifikasi dengan SVM

Penelitian ini menggunakan algoritma SVM yang emanfaatkan hyperplane untuk pemodelan. Berikut formulasi dari SVM (Cortes dan Vapnik, 1995; Akay, 2009):

$$\text{maximize}_{\beta}\left[\sum_{i=1}^{n}\beta_i - \frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j y_i y_j\, K(x_i x_j)\right]$$
$$\text{subject to } \sum_{i=1}^{n}\beta_i y_j = 0, 0 \le \forall\beta_i \le L \quad (8)$$

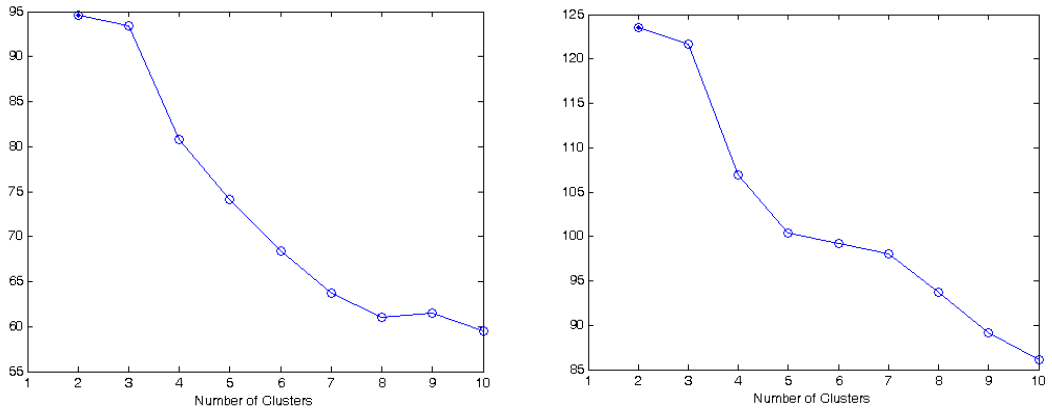
Dengan x : *training vector*, y : label dari *training vector*, β : vector parameter dari *hyperplane classifier*, K : *Kernel function*, menghitung jarak antara *training vector* x_i dan x_j, L : parameter pinalti untuk mengontrol jumlah yang salah diklasifikasi, semakin besar L, semakin akurat hasil klasifikasi.

3. HASIL EKSPERIMEN

Dari *dataset liver*, *breast cancer*, dan *heart disease* dilakukan pemisahan berdasarkan kelas pada setiap *dataset*-nya yang sebelumnya sudah dilakukan praproses. Data yang telah dipisah selanjutnya akan dievaluasi jumlah *cluster* yang optimum dengan menggunakan kriteria Calinski-Harabasz. Setiap kelas pada setiap dataset dikelompokkan dengan algoritma K-Means sejumlah optimum cluster yang telah diperoleh sebelumnya kemudian di rekonstruksi. Hasil rekonstruksi ciri ini digunakan dalam klasifikasi yakni pelatihan dan pengujian dengan algoritma SVM.

2.7 Ekstraksi dan Rekonstruksi Ciri

Pada *dataset liver*, *record* dibagi menjadi dua yakni kelas *liver* dan *nonliver*. Masing-masing data perkelas ini dikelompokkan dengan K-Means, jumlah *cluster* diperoleh dengan mengevaluasi Kriteria Calinski-Harabasz (CH). Evaluasi jumlah *cluster* dengan kriteria ini dilakukan pada range *cluster* 2 hingga 10. Gambar 3 di bawah menunjukkan jumlah *cluster* yang optimum untuk tiap kelas *liver* dan *nonliver* pada *dataset liver*.

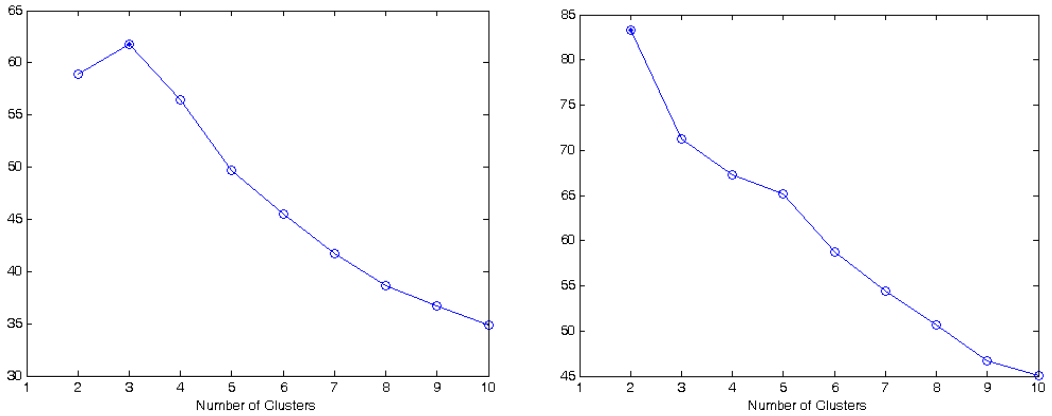


(a) (b)

Gambar 3. Cluster Optimum pada Dataset Liver (a) Kelas Liver (b) Kelas Nonliver

Kriteria Calinski Harabasz yang menunjukkan jumlah *cluster* yang optimum adalah kriteria dengan nilai tertinggi. Pada Gambar 3, data kelas *liver* dan *nonliver* dapat terpisah secara baik bila dikelompokkan ke dalam dua *cluster* yang ditunjukkan dengan nilai Calinski Harabasz maksimum. Pemisahan pada kelas *liver* ditunjukkan pada Gambar 3 (a) yang memvisualisasikan jumlah *cluster* optimum dengan dua *cluster* serta kelas non*liver* terpisah dengan baik menurut kriteria Calinski Harabasz pada jumlah *cluster* sebanyak 2 yang ditunjukkan pada Gambar 3 (b), dimana puncak grafik berada pada number of *cluster* sejumlah 2.

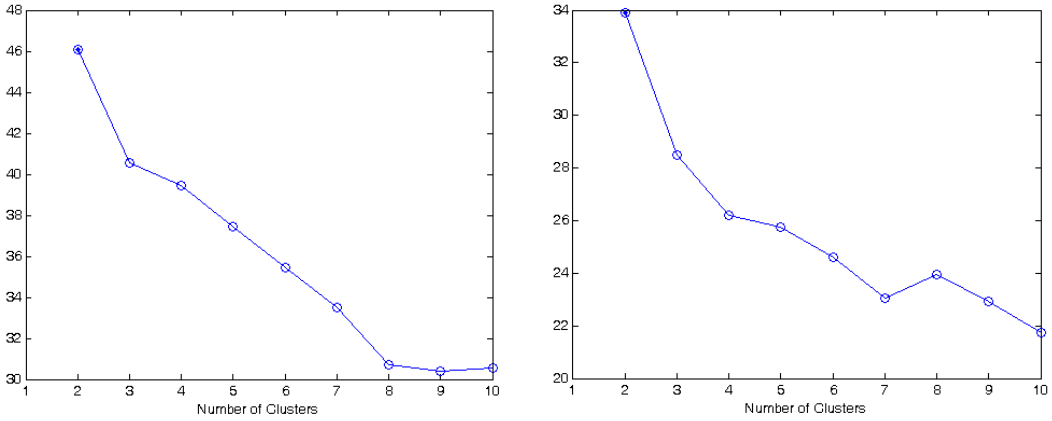
Pada *dataset breast cancer*, record dibagi menjadi dua yakni kelas kanker jinak (*benign*) dan kanker ganas (*malignant*). Masing-masing data perkelas ini dikelompokkan dengan kmeans, jumlah *cluster* diperoleh dengan mengevaluasi Kriteria Calinski-Harabasz. Evaluasi jumlah *cluster* dengan criteria ini dilakukan pada range *cluster* 2 hingga 10. Gambar 4 di bawah menunjukkan jumlah *cluster* yang optimum untuk tiap kelas *benign* dan *malignant* pada *dataset breast cancer*.



(a) (b)

Gambar 4. *Cluster* Optimum pada *Dataset Breast Cancer* (a) Kelas Malignant (b) Kelas Benign

Gambar 4 menujukkan indeks Calinski Harabasz pada kelas malignant dan benign. Pemisahan pada kelas malignant ditunjukkan pada Gambar 4 (a) yang memvisualisasikan jumlah *cluster* optimum dengan indeks tertinggi pada pengelompokan ke dalam tiga *cluster*, serta kelas benign terpisah dengan baik menurut kriteria Calinski Harabasz pada jumlah *cluster* sebanyak dua yang ditunjukkan pada Gambar 4 (b), dimana puncak grafik berada pada *number of cluster* sejumlah dua.



(a) (b)

Gambar 5. *Cluster* Optimum pada *Heart Disease Dataset* (a) Kelas Sehat (b) Kelas Sakit

Gambar 5 menunjukkan evaluasi pemisahan *cluster* dari *dataset* penyakit *heart disease* yang dikelompokkan menjadi kelas sehat dan sakit. Berdasarkan evaluasi Calinski-Harabasz, kelas sehat optimum di kelompokkan ke dalam 2 *cluster* seperti yang ditunjukkan pada Gambar 5 (a), sedangkan kelas sakit optimum di kelompokkan ke dalam 2 *cluster* pula yang divisualisasikan pada Gambar 5 (b) sebagai nilai maksimum.

Setelah diperoleh jumlah *cluster optimum* untuk masing-masing kelas pada setiap *dataset* dan pengelompokan dengan K-Means sejumlah *cluster optimum* tersebut, dilakukan rekonstruksi ciri, yakni membentuk ciri baru berdasarkan hasil *clustering* dengan formula [6] dan [7]. Hasil perhitungan rekonstruksi ciri

dengan *fuzzy membership function* dan pemnbentukan pola ini menghasilkan ciri baru berupa ciri abstrak dari setiap *dataset* dengan jumlah ciri yang telah tereduksi.

Rekonstruksi ciri ini memproses data *liver* dari sepuluh ciri menjadi empat ciri baru yang berasal dari hasil pengelompokan kelas *liver* dan *nonliver* masing-masing dua kelompok. Data *breast cancer* dari tiga puluh ciri menjadi lima ciri, serta data penyakit *heart disease* dari tiga belas ciri menjadi empat ciri baru. *Dataset* dengan ciri-ciri baru tersebut selanjutnya diproses pada fase latihan untuk menghasilkan *classifier* menggunakan algoritma SVM yang kemudian akan divalidasi dengan data uji.

2.8 Klasifikasi dengan SVM

Eksperimen dilakukan dengan *5-fold cross validation* pada setiap *dataset* dan evaluasi performa dari metode ini dilakukan dengan menghitung akurasi pengujian serta waktu komputasi dari proses latihan, dimana akurasi merupakan persentase dari hasil klasifikasi yang diprediksi benar, dibandingkan dengan klasifikasi actual sebagaimana dapat dilihat pada formula 9.

$$akurasi = \frac{prediksi}{aktual} \times 100\% \quad (9)$$

Pengujian dilakukan dengan melakukan pengulangan eksperimen sebanyak sepuluh kali dan mengambil rata-rata performa setiap eksperimen, yakni rata-rata nilai akurasi pengujian dan waktu komputasi latihan pada sepuluh percobaan untuk menjamin konsistensi hasil pada setiap *dataset liver, breast cancer* dan *heart disease*. Untuk membandingkan performa Klasifikasi SVM yang melalui tahap ekstraksi ciri dengan K-Means dan tanpa K-Means, dapat dilihat pada Tabel 1.

Tabel 1. Perbandingan akurasi SVM dengan ekstraksi K-Means dan tanpa ekstraksi K-Means

	K-Means & SVM (%)	*SVM (%)*
Liver	89.87	90.60
Breast Cancer	95.70	96.17
Heart Disease	83.38	84.88

Tabel 1 menunjukkan hasil klasifikasi dengan SVM dengan proses ekstraksi ciri terlebih dahulu dengan K-Means dan tanpa ekstraksi ciri. Dari hasil tersebut terlihat bahwa dengan ektraksi ciri K-Means, akurasinya menjadi lebih kecil. Untuk *dataset liver*, akurasi menurun 0.73%, *breast cancer* 0.46%, dan *heart disease* 1.5%. Hal ini menunjukkan hasil yang berbeda dengan penelitian Zheng et.al. (2014) yang menunjukkan peningkatan akurasi.

Tabel 2. Perbandingan waktu komputasi SVM dengan ekstraksi K-Means dan tanpa ekstraksi K-Means

	K-Means & SVM (detik)	*SVM (detik)*
Liver	0.67	1.04
Breast Cancer	0.33	0.68
Heart Disease	0.38	1.20

Dari sisi waktu komputasi saat latihan, hasil eksperimen dapat dilihat pada Tabel 2. Terlihat perbedaan yang cukup signifikan pada penggunaan ekstraksi ciri dengan K-Means, yakni penurunan waktu komputasi. Pada *dataset liver*, latihan tanpa ekstraksi memerlukan waktu rata-rata 1.04 detik, sedangkan dengan ekstraksi K-Means memerlukan waktu rata-rata 0.67 detik. pada *dataset breast cancer*, penurunan waktu komputasi mencapai separuh dari waktu awal tanpa ekstraksi ciri, sedang pada *dataset heart disease* penurunan sangat signifikan dari rata-rata waktu komputasi 1.2 detik menjadi 0.38 detik.

Tabel 3. Penurunan akurasi dan peningkatan komputasi dengan ekstraksi K-Means

	Penurunan Jumlah Ciri	*Selisih Akurasi (%)*	*Selisih Kecepatan Waktu Komputasi (detik)*
Liver	10 → 4	- 0.73	0.36
Breast cancer	30 → 5	- 0.46	0.35
Heart disease	13 → 4	- 1.50	0.82

Persentase perubahan akurasi serta waktu komputasi dari eksperimen dapat dilihat pada Tabel 3. Setelah dilakukan ekstraksi dan rekonstruksi yang memanfaatkan K-Means serta fuzzy membership function, penurunan jumlah ciri cukup signifikan, yakni *liver* sebelumnya memiliki 10 ciri menjadi 4 ciri dengan waktu komputasi saat latihan lebih cepat hingga 0.36 detik disertai penurunan akurasi sebesar 0.73%. Sedangkan pada *breast cancer* mengalami penurunan jumlah ciri dari 30 menjadi 5 disertai penurunan akurasi sebesar 0.46% dan peningkatan waktu komputasi lebih cepat 0.35 detik. Penurunan akurasi terbesar pada data *heart disease* dengan selisih 1.5%, disertai peningkatan kecepatan waktu komputasi sebesar 0.82 detik. Sehingga, proses ekstraksi dan ciri dengan K-

Means cukup mempengaruhi performa dari klasifikasi, yakni meningkatkan performa komputasi dengan sedikit penurunan akurasi.

4. KESIMPULAN DAN SARAN

Pada penelitian ini, dilakukan evaluasi penggunaan algoritma *clustering* K-Means sebagai ekstraktor ciri pada klasifikasi data *liver, breast cancer*, dan *heart disease* dengan SVM. Metode ini mengekstrak ciri berdasarkan pola *cluster* dari setiap kelas dan merekonstruksi kembali data berdasarkan pola yang telah diekstrak serta menghasilkan ciri abstrak yang akan digunakan dalam latihan untuk mendapatkan klasifikasi.

Penggunaan K-Means sebagai ekstraktor tidak terlalu banyak menurunkan akurasi, tapi signifikan dalam menurunkan waktu komputasi terutama pada data dengan dimensi ciri yang besar, sehingga dapat dipertimbangkan sebagai suatu metode yang mampu meningkatkan performa komputasi terutama klasifikasi dengan SVM.

Semakin hari, data akan semakin meningkat jumlahnya dan semakin banyak sample data, dan dengan perkembangan teknologi akan semakin besar dimensi feature yang diperoleh. Oleh karena itu, metode ekstraksi ciri akan selalu perlu untuk dieksplorasi lebih mendalam untuk mengatasi berbagai masalah yang muncul.

DAFTAR PUSTAKA

[1] A. Janecek, W. N. W. Gansterer, M. Demel, and G. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy.," *Fsdm*, vol. 4, pp. 90–105, 2008.

[2] C. Lee and D. Landgrebe, "FEATURE EXTRACTION AND CLASSIFICATION ALGORITHMS FOR HIGH DIMENSIONAL DATA," 1993.

[3] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 145–171, Jan. 2015.

[4] H. Cen, R. Lu, Q. Zhu, and F. Mendoza, "Nondestructive detection of chilling injury in cucumber fruit using hyperspectral imaging with feature selection and supervised classification," *Postharvest Biol. Technol.*, vol. 111, pp. 352–361, Jan. 2016.

[5] A. T. Azar and A. E. Hassanien, "Dimensionality reduction of medical big data using neural-fuzzy classifier," *Soft Comput.*, vol. 19, no. 4, pp. 1115–1127, Apr. 2015.

[6] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.," *Adv. Bioinformatics*, vol. 2015, p. 198363, Jun. 2015.

[7] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, and G. Giacinto, "Novel Feature Extraction, Selection and Fusion for Effective Malware Family Classification," in *Proceedings of the Sixth ACM on Conference on Data and Application Security and Privacy - CODASPY '16*, 2016, pp. 183–194.

[8] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, Mar. 2014.

[9] N. Chamidah and I. Wasito, "Fetal state classification from cardiotocography based on feature extraction using hybrid K-Means and support vector machine," in *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2015, pp. 37–41.

[10] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[11] D. Dheeru and E. K. Taniskidou, "UCI machine learning repository," *University of California, Irvine, School of Information and Computer Sciences*, 2017. .

[12] N. Chamidah, Wiharto, and U. Salamah, "Pengaruh Normalisasi Data pada Jaringan Syaraf Tiruan Backpropagasi Gradient Descent Adaptive Gain (BPGDAG) untuk Klasifikasi," *ITSMART J. Teknol. dan Inf.*, vol. 1, no. 1, pp. 28–33, Sep. 2012.

[13] S. Lukasik, P. A. Kowalski, M. Charytanowicz, and P. Kulczycki, "Clustering using flower pollination algorithm and Calinski-Harabasz index," *2016 IEEE Congr. Evol. Comput. CEC 2016*, no. 1, pp. 2724–2728, 2016.

[14] B. Halpin, Halpin, and Brendan, "CALINSKI: Stata module to compute Calinski-Harabasz cluster stopping index from distance matrix," Jun. 2016.

[15] T. Caliński and J. Harabasz, "A Dendrite Method Foe Cluster Analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.