Gede Suwardika. (2017). Pengklasifikasian Pada Data *Echocardiogram* Dengan Menggunakan *Support Vector Machine* dan Analisis Diskriminan. International Journal of Natural Science and Engineering. Vol.1 (1) pp. 1-7.

Pengklasifikasian Pada Data *Echocardiogram* Dengan Menggunakan *Support Vector Machine* dan Analisis Diskriminan

Gede Suwardika [1,*]

[1] *Unit Program Belajar Jarak Jauh Universitas Terbuka Denpasar*

Abstrak

Echocardiogram (seringkali disebut "echo") adalah garis luar grafik dari gerakan jantung. Selama tes ini, gelombang-gelombang suara frekwensi tinggi, disebut ultrasound, menyediakan gambar-gambar dari klep-klep dan kamar-kamar jantung. Dalam penelitian ini dilakukan tes terhadap 132 pasien dengan respon meninggal atau hidup. Hasil ketepatan klasifikasi antara data training dengan data testing dengan analisis diskriminan adalah 96% sedangkan dengan menggunakan SVM diperoleh sebesar 88%. Pengelompokan dengan menggunakan K-Means dan Kernel K-Means menghasilkan ketepatan pengelompokan yang sama persis. Ini menunjukkan bahwa data echocardiogram memiliki pengelompokan yang baik. Kemudian hasil pengelompokan pada K-Means dibandingkan dengan data aktual yang diklasifikasikan dengan menggunakan diskriminan, SVM dan CART dimana dihasilkan bahwa data hasil dari K-Means memiliki ketepatan klasifikasi yang lebih baik dibandingkan dengan hasil klasifikasi pada data aktual.

Keywords:
Analisis Diskriminan, SVM, K-Means, CART.

PENDAHULUAN

Echocardiogram (seringkali disebut "echo") adalah garis luar grafik dari gerakan jantung. Selama tes ini, gelombang-gelombang suara frekwensi tinggi, disebut ultrasound, menyediakan gambar-gambar dari klep-klep dan kamar-kamar jantung. Dalam penelitian ini dilakukan tes terhadap 132 pasien dengan respon meninggal atau hidup Edler I. 2004).

Pattern Recognition merupakan salah satu bidang dalam komputer sains, yang memetakan suatu data ke dalam konsep tertentu yang telah didefinisikan sebelumnya. Konsep tertentu ini disebut class atau category. Aplikasi pattern recognition sangat luas, di antaranya mengenali suara dalam sistem sekuriti, membaca huruf dalam OCR, mengklasifikasikan penyakit secara otomatis berdasarkan hasil diagnosa kondisi medis pasien dan sebagainya. Berbagai metode dikenal dalam *pattern recognition*, seperti linear discrimination analysis, hidden markov model hingga metode kecerdasan buatan seperti artificial neural network. Salah satu metode yang akhir-akhir ini banyak mendapat perhatian sebagai state of the art dalam pattern recognition adalah Support Vector Machinev (SVM)

Penggunaan sistem klasifikasi dalam diagnosis medis meningkat secara bertahap. Machine learning telah banyak digunakan dalam bidang medis untuk menganalisa dataset medis (Munawarah, Raudlatul, 2016). Salah satu metode machine learning adalah Support Vector Machine (SVM). Dalam kasus ini, kita akan mencoba menganalisa dengan cara menggunakan *Support Vector Machine* (SVM). Kemudian dilakukan pengklasifikasian kedua dengan menggunakan analisis diskriminan untuk membandingkan seberapa akurat pengklasifikasian dengan 2 metode tersebut.

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (feature space) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik (Raudlatul Munawarah1, Oni Soesanto2, M. Reza Faisal, 2016). SVM adalah sistem *learning* yang menggunakan sebuah ruang hipotesis fungsi linier dalam ruang fitur berdimensi tinggi (Vladimir N. Vapnik,1995), dilatih dengan menggunakan sebuah algoritma pembelajar dari teori optimasi yang mengimplementasikan sebuah bias *learning* yang diturunkan dari teori *learning* statistika yang mana nantinya akan dibandingkan dengan klasifikasi dengan analisis diskriminan. Menurut Ramana, dkk, 2011) SVM adalah salah satu teknik yang relatif baru dibandingkan

* Corresponding author.
E-mail Addresses: isuwardika@ecampus.ut.ac.id (Gede Suwardika)

dengan teknik lain, tetapi memiliki performansi yang lebih baik di berbagai bidang aplikasi seperti bioinformatics, pengenalan tulisan tangan, klasifikasi teks dan lain sebagainya. Support vector machine (SVM) sebagai salah satu metode dari data mining terbukti memiliki tingkat akurasi yang tinggi dalam mengklasifikasikan pola-pola paket data jaringan (Jacobus, 2013).

Suatu Support Vector Machine (SVM) melakukan klasifikasi dengan membangun sebuah hyperplane N-dimensi yang optimal memisahkan data menjadi dua kategori. Model SVM terkait erat dengan jaringan saraf. Bahkan, model SVM menggunakan fungsi kernel sigmoid adalah setara dengan dua lapisan, perceptron neural network. Model Support Vector Machine (SVM) adalah saudara dekat dengan multilayer perceptron klasik jaringan saraf.

Menggunakan fungsi kernel, yang SVM adalah metode pelatihan alternatif untuk polinomial, fungsi dasar radial dan multi-lapisan pengklasifikasi perceptron di mana bobot jaringan ditemukan dengan memecahkan masalah pemrograman kuadrat dengan kendala linear, bukan dengan memecahkan non-cembung, minimisasi masalah tidak dibatasi seperti dalam pelatihan jaringan saraf standar. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah kelas pada input space. Pattern yang merupakan anggota dari dua buah kelas : +1 dan -1 dan berbagi alternative garis pemisah (discrimination boundaries). Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing kelas. Pattern yang paling dekat ini disebut sebagai support vector. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pembelajaran pada SVM (Christianini, Nello dan John S. Taylor. 2000).

Dalam bahasa sastra SVM, variabel prediktor disebut atribut, dan atribut berubah yang digunakan untuk menentukan hyperplane disebut fitur. Tugas memilih representasi yang paling cocok adalah dikenal sebagai seleksi fitur. Satu set fitur yang menggambarkan satu kasus (misalnya, deretan nilai-nilai prediktor) disebut vektor. Jadi tujuan dari pemodelan SVM adalah untuk menemukan hyperplane optimal yang memisahkan cluster dari vektor sedemikian rupa sehingga kasus dengan satu kategori dari variabel target pada satu sisi pesawat dan kasus-kasus dengan kategori yang lain pada ukuran lain dari pesawat. Vektor-vektor di dekat hyperplane adalah vektor-vektor dukungan.

METODE PENELITIAN

Data yang digunakan *echocardiogram* yang berasal dari UCI *Machine Learning Repository*.
Variabel-variabel yang digunakan:

Variabel bebas (X):

1. Survival: Jumlah bulan pasien selamat (selamat, jika pasien masih hidup). Karena semua pasien mengalami serangan jantung pada waktu yang berbeda, adalah kemungkinan bahwa beberapa pasien bertahan kurang dari satu tahun tetapi mereka masih hidup. Periksa kedua variabel untuk mengkonfirmasi ini. Pasien tersebut tidak dapat digunakan untuk tugas prediksi tersebut di atas.
2. Yang masih hidup. Merupakan sebuah variabel biner.
   0 = mati pada akhir periode survival
   1 = berarti masih hidup
3. Usia (dalam tahun) ketika serangan jantung terjadi.
4. Efusi perikardial. Efusi perikardial adalah cairan sekitar jantung.
   0 = tidak ada cairan
   1 = cairan
5. Fractional Shortening: ukuran contracility sekitar jantung, angka yang lebih rendah semakin tidak normal.
6. Epss: E-titik pemisahan septum, ukuran lain kontraktilitas. Angka yang lebih besar semakin abnormal.
7. Lvdd: Dimensi ventrikel akhir diastolik kiri. Ini adalah ukuran dari ukuran jantung pada akhir diastole. Hati yang besar cenderung sakit.
8. Skor dinding gerak: Ukuran bagaimana segmen dari kiri ventrikel bergerak
9. Indeks dinding gerak: sama dengan skor dinding gerak dibagi dengan jumlah segmen yang terlihat. Biasanya segmen 12-13 dilihat di echocardiogram. Gunakan variabel ini bukan dari skor gerakan dinding.

Variabel respon (Y):

Alive at 1: Nilai Boolean. Berasal dari dua atribut pertama.
0 = berarti pasien mati setelah 1 tahun atau memiliki telah diikuti selama kurang dari 1 tahun.
1 = berarti pasien masih hidup pada 1 tahun.

Langkah-langkah yang dilakukan dalam penelitian ini adalah sebagai berikut: a) Mengevaluasi data *echocardiogram* apakah terdapat *missing value*. Kemudian setelah diketahui *missing value* yang besar pada variable, maka dilakukan penghapusan pada variable tersebut, b) Mengevaluasi data *missing value* berdasarkan observasi. Jika *missing value* memiliki persentase yang besar, maka dihilangkan, c) Untuk variable dan observasi yang memiliki *missing value* tetapi tidak dalam jumlah besar, kekosongan nilai dapat diisi dengan mean dari tiap-tiap variabel, d) Mengklasifikasikan data dengan menggunakan analisis diskriminan, e) Mengklasifikasikan data menggunakan SVM dengan bantuan *software Matlab* setelah sebelumnya membagi data menjadi 100 data *training* dan 25 data sebagai *testing*, f) Pembandingan ketepatan klasifikasi antara analisis diskriminan dan SVM, g) Mengklasifikasikan dengan menggunakan K-Means dan Kernel K-Means kemudian menentukan hasil prediksi terbaik yang mendekati data aktual, h) Data aktual diklasifikasikan menggunakan analisis diskriminan, SVM dan CART kemudian membandingkan hasilnya, i) Hasil prediksi terbaik pada langkah 7 diklasifikasikan menggunakan analisis diskriminan, SVM dan CART kemudian membandingkan hasilnya.

ANALISIS DAN PEMBAHASAN

Langkah pertama yang dilakukan adalah *Preprocessing* Data *Missing Value*. Pada tahap ini akan dilakukan pengevaluasian terhadap banyaknya *missing value*. Variabel Y memiliki banyak *missing value* yaitu sebesar 58%, sehingga variabel Y dihilangkan. Karena variable Y merupakan variabel respon yang berasal dari 2 variabel pertama, maka variabel X_2 yang bertindak sebagai variable respon.

Berdasarkan *case*/ observasi yang memiliki presentase *missing value* yang besar juga dihilangkan, sehingga perlu dilakukan penghapusan data berdasarkan *case*-nya yaitu case ke 85, 49, 46, 28, 33, 29, 50. Pada beberapa variabel dan observasi yang memiliki *missing value* dengan persentase kecil, maka kekosongan nilai dapat diisi dengan mean yang diperoleh dari masing-masing variable seperti pada Tabel 1.

Tabel 1. Nilai Mean Yang Dapat Diisikan Pada *Missing value*

		v1	v2	v3	v4	v5	v6	v7	v8	v9
N	Valid	124	125	121	125	123	115	121	123	124
	Missing	1	0	4	0	2	10	4	2	1
Mean		22.6898	.3120	62.7812	.1840	.2178	12.2111	4.7632	14.4478	1.3509
Sum		2813.53	39.00	7596.53	23.00	26.79	1404.28	576.34	1777.08	167.51

Selanjutnya dilakukan Analisis Diskriminan Untuk melakukan analisis diskriminan terhadap sebuah data, perlu dilakukan uji asumsi terlebih dahulu yang meliputi uji normal multivariate dan uji kehomogenan matriks varian kovarian. Pada analisis diskriminan digunakan 100 data sebagai *training* dan 25 sisanya sebagai data *testing*. Terdapat asumsi yang harus dipenuhi dalam analisis diskriminan yaitu bahwa data harus memenuhi asumsi berdistribusi normal multivariate dan matrik varian kovarian antara kategori meninggal dan kategori hidup homogen.

Uji normal multivariat data digunakan untuk mengetahui apakah data berdistribusi normal multivariat atau tidak. Uji normal multivariat data menggunakan *macro minitab* untuk 7 variabel yaitu variabel v_1 sampai variabel v_7 dan hasilnya adalah sebagai berikut:

Hipotesis :

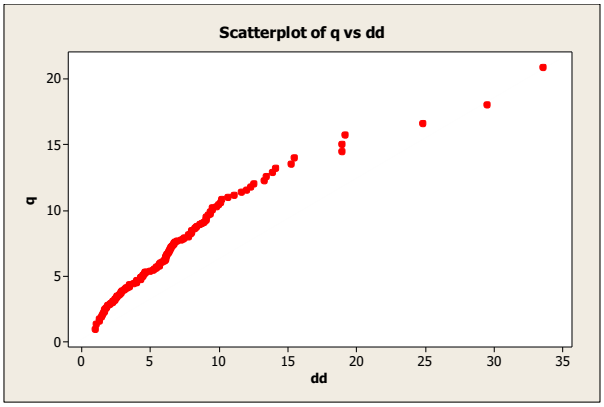
H_0 : Data *echocardiogram* berdistribusi normal multivariate.

H_1 : Data *echocardiogram* tidak berdistribusi normal multivariate.

$\alpha = 0.05$

Daerah Penolakan : jika $t > \alpha$, maka gagal tolak H_0.

Hasil pengujian data didapatkan nilai t sebesar 0.560000 maka distribusi data adalah multinormal. Kesimpulannya adalah data *echocardiogram* berdistribusi normal multivariate.



Gambar 1. *Scatterplot* Data *Echocardiogram*

Uji Kehomogenan Matrik Varian Kovarian antara Kategori 0 dan Kategori 1 dilakukan menggunakan *SPSS* yaitu dengan menggunakan statistik Box-M dan hasilnya sebagai berikut:

Hipotesis :

$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ dimana *k* = observasi sampai ke-*k.*

$H_1 : \Sigma_i \neq \Sigma_j$ untuk $i \neq j$ dimana i,j = 1,2,...,n.

$\alpha = 0.05$

Daerah penolakan : jika nilai signifikan < α, maka gagal tolak H_0.

Hasil pengujian data tersebut didapatkan nilai p_value sebesar 0.030 maka dapat disimpulkan gagal tolak H_0 karena nilai signifikan < α yang artinya matriks varians-kovarians antara kategori meninggal dan kategori hidup homogen.

Karena hasil 2 uji asumsi di atas menunjukkan bahwa data memenuhi asumsi berdistribusi normal multivariate dan matrik varian kovarian antara kategori meninggal dan kategori hidup homogen, maka data yang meliputi 100 data *training* dan 25 data *testing* dapat dianalisis diskriminan dengan menggunakan *SPSS*.

Dari hasil *Group Statistics* pertama, diperoleh informasi bahwa untuk kategori 0 yaitu meninggal terdapat 4 variabel sebagai variabel diskriminator karena memiliki nilai standart deviasi yang kurang dari total standart deviasi. Keempat variabel tersebut adalah v_1, v_3, v_4 dan v_5. Sedangkan sisanya adalah variabel yang kurang baik sebagai variabel diskriminator karena memiliki nilai standart deviasi yang lebih besar dari pada total standart deviasi.

Untuk kategori 1 yaitu kategori hidup diperoleh informasi bahwa juga terdapat 4 variabel sebagai variabel diskriminator karena memiliki nilai standart deviasi yang kurang dari total standart deviasi. Keempat variabel tersebut adalah v_1, v_4, v_6 dan v_7. Sedangkan sisanya adalah variabel yang kurang baik sebagai variabel diskriminator karena memiliki nilai standart deviasi yang lebih besar dari pada total standart deviasi.

Selanjutnya mencari fungsi diskriminan. Untuk dapat memperoleh fungsi diskriminan maka digunakan tabel *canonical discriminant function* dan didapatkan bahwa rumus untuk fungsi diskriminannya adalah: $D = 0.894v_1-0.268v_2+0.363v_3$. Fungsi diskriminan ini dapat digunakan untuk variabel *still alive* yaitu pasien meninggal atau hidup (v_8). Pasien tersebut dan dihitung *score*-nya berdasarkan fungsi diskriminan yang diperoleh.

Dari hasil output diperoleh informasi bahwa pada kategori 0 yang diprediksi tepat berada pada kategori 0 sebanyak 3 data sedangkan 1 sisanya tidak tepat berada pada kategorinya. Kemudian pada kategori 1 yang diprediksi tepat pada kategori 1 seluruh data yaitu sebanyak 21 data. Persentase ketepatan data berada pada kategori yang benar adalah 96%.

SVM adalah suatu teknik yang relatif baru untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. Dalam hal ini data yang ingin diklasifikasikan adalah data *echocardiogram.* Dalam pengklasifikasian SVM ini ingin diketahui variabel y prediksi berdasarkan data *training* dan data *testing*, sehingga nantinya diketahui ketepatan y prediksi terhadap variabel y yang sebenarnya. Dengan bantuan program *Matlab*, didapatkan hasil seperti Tabel 2 berikut ini:

Tabel 2. Hasil Prediksi Y Dan Y *Testing*

Data *Training*	Data *Testing*	Hasil Prediksi		Ketepatan Klasifikasi
		Sesuai	Tidak sesuai	
80	45	38	7	84%
90	35	30	5	85%
100	25	22	3	88%

Jadi ketepatan klasifikasi dengan SVM antara y prediksi dengan y *testing* yang terbaik adalah sebesar 88% dengan *training* sebanyak 100 dan 25 sebagai *testing*. Selanjutnya pengelompokan pada data aktual *echocardiogram* dengan menggunakan K-Means menghasilkan pengelompokan seperti pada Tabel 3.

Tabel 3. Hasil Pengelompokan Menggunakan K-Means

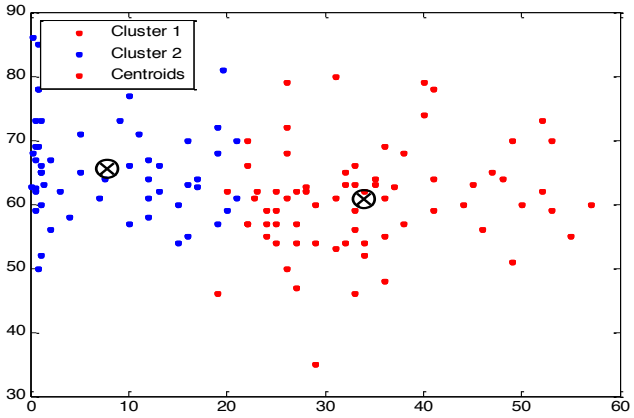
Kelas	Total	Hasil Klasifikasi
-1	71	86
1	54	39

Dari hasil pengelompokan menggunakan K-Means, dapat dilihat bahwa sebanyak 71 data dikelompokkan dalam kelas pertama, sedangkan sisanya sebanyak 54 data dikelompokkan dalam kelas kedua. Hasil yang didapatkan dari K-Means akan dibandingkan dengan kernel K-Means yang kemudian dibandingkan hasilnya untuk mendapatkan hasil pengelompokan terbaik yang mendekati data aktual. Dengan menggunakan kernel 'rbf' (*Radial Basis Function*) dan 'poly' menghasilkan *error* terkecil didapatkan hasil klasifikasi seperti pada Tabel 4.

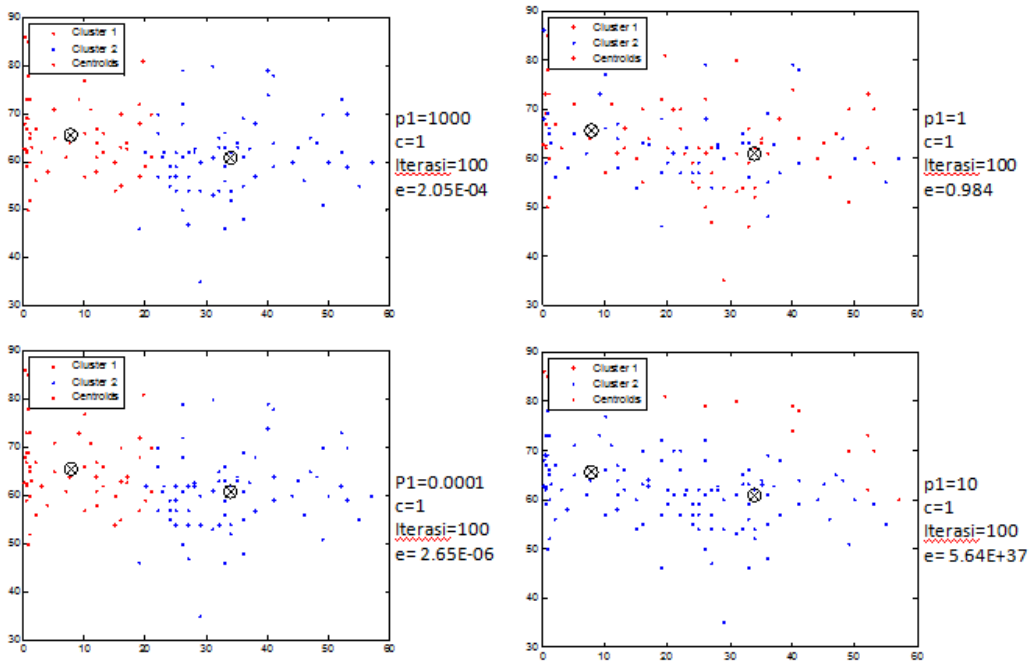
Tabel 4. Hasil Pengelompokan Menggunakan Kernel K-Means

Kelas	Total	Hasil Klasifikasi
-1	71	86
1	54	39

Dari Table 3. dan Table 4. dapat dilihat bahwa dihasilkan pengelompokan yang sama persis. Gambar 2. menunjukkan plot hasil pengelompokan dengan menggunakan kedua metode. Pada plot kernel dibandingkan dengan hasil kernel dengan *error* yang besar sebagai pembanding.



K-Means



Kernel K-Means

Gambar 2**.** Plot Hasil Pengelompokan Dengan K-Means dan Kernel K-Means

Berdasarkan hasil yang didapatkan menggunakan kedua metode, maka diambil salah satu yang terbaik yaitu hasil dari K-Means.

Tahap selanjutnya adalah melakukan klasifikasi data aktual dan hasil k-means menggunakan analisis diskriminan, svm dan cart. Pengklasifikasian pada data aktual berikut ini tanpa membagi data menjadi data *testing* dan data *training* dilakukan untuk mengetahui bagaimana hasil ketepatan klasifikasi dengan menggunakan analisis diskriminan, SVM dan CART. Hasil pengklasifikasian dibandingkan dengan klasifikasi dengan menggunakan data hasil pengelompokan dengan menggunakan K-Means. Tabel 5. merupakan perbandingan hasil dengan menggunakan kedua data dengan analisis diskriminan, SVM dan CART.

Tabel 5. Perbandingan Ketepatan Klasifikasi

Y	**Metode**		
	Diskriminan	**SVM**	**CART**
Data Aktual	83.20%	92.80%	93.60%
Hasil K-Means	95.20%	100%	99.20%

Pengklasifikasian antara kedua data dengan ketiga metode menghasilkan ketepatan klasifikasi dengan nilai yang besar. Pada data aktual, ketepatan klasifikasi terbesar adalah dengan menggunakan metode CART yaitu sebesar 93.6%, sedangkan pada data hasil K-Means dihasil ketepatan sebesar 100% dengan menggunakan SVM.

Berdasarkan Table 5. dapat dilihat bahwa ketepatan klasifikasi kedua data menggunakan ketiga metode menghasilkan ketepatan yang lebih besar dengan menggunakan data hasil dari K-Means. Hal ini dikarenakan data kedua sudah merupakan hasil dari pengelompokan dengan metode K-Means, sehingga pengklasifikasian-nya akan lebih baik dibandingkan dengan data aktual.

KESIMPULAN

Kesimpulan yang dapat dibuat berdasarkan hasil klasifikasi yang telah dilakukan adalah 1) Hasil ketepatan klasifikasi antara y prediksi dengan y *testing* dengan analisis diskriminan adalah 96% sedangkan dengan menggunakan SVM diperoleh sebesar 88%. Dalam hal ini klasifikasi menggunakan analisis diskriminan menghasilkan ketepatan yang lebih besar dibandingkan dengan menggunakan metode SVM, 2) Pengklasifikasian dengan menggunakan K-Means dan Kernel K-Means menghasilkan ketepatan klasifikasi yang sama persis. Ini menunjukkan bahwa data *echocardiogram* memiliki pengelompokan yang baik, 3) Pada data aktual, ketepatan klasifikasi terbesar adalah dengan

menggunakan metode CART yaitu sebesar 93.6%, sedangkan pada data hasil K-Means dihasil ketepatan sebesar 100% dengan menggunakan SVM. Secara global dapat dilihat bahwa data hasil dari K-Means menghasilkan ketepatan klasifikasi yang lebih baik dibandingkan dengan hasil klasifikasi pada data aktual.

DAFTAR PUSTAKA

Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone (1984).Classication and Regression Trees. Belmont, California: Wadsworth.

D.C. Montgomery. 1991. *Design and Analysis of Experiments*, Third Edition. John Wiley & Sons.

Christianini, Nello dan John S. Taylor. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press

Edler I. 2004. *The History of Echocardiography.* Departemen Kardiologi, Universitas Hospital, Lund, Sweden. http://www.ncbi.nlm.nih.gov diakses pada tanggal 5 januari 2012.

Jacobus, Agustinus, Edi Winarko. 2013. Penerapan Metode Support Vector Machine pada Sistem Deteksi Intrusi secara Real-time. Berkala MIPA, 23 (2).

J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

Munawarah, Raudlatul , Oni Soesanto, M. Reza Faisal. 2016. Penerapan Metode Support Vector Machine Pada Diagnosa Hepatitis. Kumpulan jurnaL Ilmu Komputer (KLIK) Volume 04, No.01.

N. Cristianini, J. Shawe-Taylor, An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge University Press, New York, NY, USA (2000)

Ramana, B. V., Babu, S. P., & Venkateswarlu, N. B. 2011. A Critical Study Of Selected Classification Algorithms For Liver Disease Diagnosis. International Journal Of Database Management Systems , Vol. 3 (2), Hal. 101-114.

Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.