Optimasi *Support Vector Machine* untuk Memprediksi Adanya Mutasi pada DNA *Hepatitis C Virus*

Berlian Al Kindhi[1,5], Tri Arief Sardjono[2,3], Mauridhi Hery Purnomo[2,4]

***Abstract*— Hepatitis C Virus (HCV) is a virus which capable of infecting RNA that can lead to changes in the DNA sequence. This change of DNA arrangement is called genetic mutation. Every mutation occurs in HCV, it will be called a new subtype. Over time, HCV subtypes increase, and will continue to grow as the HCV mutation cycle progresses faster. Therefore, a way to find a mutation in millions of sequences in the gene bank is needed. This study tested six types of Support Vector Machine (SVM) methods to determine the best SVM kernel performance in the application of HCV DNA sequence detection in isolated DNA. The tested SVM kernel was linear, quadratic, cubic, fine Gaussian, median Gaussian, and coarse Gaussian. The data set is 1000 isolated DNA consisting of 500 isolated Homo Sapiens and 500 isolated HCV. Firstly, the data set will go through the pattern search process using the Edit Levenshtein Distance method, then the result of the processing will be the variable x in SVM. The target or variable *y* on SVM is the positive or negative value of the isolated against HCV. The results show that among the six types of SVM methods being tested, the method of fine Gaussian SVM has the lowest performance of 77.4%. The SVM method is tested by performing optimizations on the determination of the hyperplane. The test results proved that the SVM method is able to analyze the presence of HCV mutations in isolated DNA with an accuracy of 99.8%.**

***Intisari*— *Hepatitis C Virus* (HCV) adalah virus yang mampu menginfeksi RNA sehingga dapat mengakibatkan perubahan susunan DNA. Perubahan susunan DNA inilah yang disebut dengan mutasi genetik. Setiap terjadi mutasi pada HCV, akan disebut dengan subtipe baru. Semakin lama, subtipe HCV semakin banyak, dan akan terus bertambah seiring dengan semakin cepatnya siklus mutasi HCV. Oleh karena itu, dibutuhkan sebuah cara yang dapat menemukan adanya mutasi pada jutaan *sequence* di dalam bank gen. Makalah ini menguji coba enam jenis metode *Support Vector Machine* (SVM) untuk mengetahui kinerja terbaik *kernel* SVM dalam penerapan deteksi *sequence* DNA HCV di dalam *isolated DNA*. *Kernel* SVM yang diuji yaitu *linear, quadratic, cubic, fine Gaussian, median Gaussian, dan coarse Gaussian*. *Data set* yang digunakan adalah 1000 *isolated DNA* yang terdiri atas 500 *isolated* Homo Sapiens dan 500 *isolated HCV*. *Data set* tersebut akan melalui proses pencarian pola terlebih dahulu menggunakan metode *Edit Levenshtein Distance*, kemudian hasil dari pengolahan tersebut akan menjadi variabel *x* pada SVM. Target atau variabel *y* pada SVM adalah nilai positif atau negatifnya *isolated DNA* tersebut terhadap HCV. Hasil penelitian menunjukkan bahwa dari keenam jenis metode SVM yang diujikan, metode *fine Gaussian SVM* memiliki kinerja paling rendah, yaitu sebesar 77.4%. Metode SVM diuji coba dengan melakukan optimasi pada penentuan *hyperplane*-nya. Hasil uji coba membuktikan bahwa metode SVM mampu menganalisis adanya mutasi HCV pada *isolated DNA* dengan menghasilkan akurasi sebesar 99,8%.**

***Kata Kunci*—SVM, Mesin Pembelajaran, *sequence DNA*, *Semantic Similarity*.**

I. PENDAHULUAN

Pengolahan data *Dioxyribo Nucleic Acid* (DNA) untuk mencari suatu pola dalam *sequence* sudah sering dilakukan [1]-[3]. Tujuan dari pencarian pola ini beragam, ada yang dipakai sebagai forensik, mutasi genetik, atau adanya suatu penyakit di dalam suatu DNA. DNA forensik biasanya dilakukan untuk mengenali suatu individu berdasarkan hubungan keluarga [4]. Pengenalan pola DNA juga dapat digunakan untuk mendeteksi adanya penyakit di dalam suatu *isolated DNA*, yang dapat ditandai dengan adanya perubahan susunan DNA [5]. Selain itu, pengenalan pola DNA dilakukan untuk memprediksi adanya mutasi genetik, yaitu perubahan suatu susunan DNA akibat terinfeksi oleh virus atau bakteri [6].

Untuk mencari pola di dalam suatu *isolated DNA*, dibutuhkan sebuah metode pengenalan, baik itu metode *string matching*, *pattern distance*, maupun *semantic similarity*. Metode *string matching* yang sering digunakan pada pengenalan pola DNA adalah Boyer More, Knuth Morris Pratt, dan *Brute Force*. Dari ketiga metode tersebut, metode Boyer More memiliki kinerja yang paling tinggi [7]. Untuk metode *pattern distance*, dapat digunakan metode Hamming, Hausdorff, dan *Edit Levenshtein Distance* [8]. Dari ketiga metode tersebut, metode *Edit Levenshtein Distance* memiliki kinerja paling baik. Sedangkan untuk *semantic similarity* dapat digunakan metode berbasis skor terhadap pola yang dibandingkan. Pada makalah ini digunakan metode *Edit Levenshtein Distance* sebagai parameter *preprocessing*, karena sudah melalui proses ujicoba pada penelitian yang berbeda.

Hasil dari pengenalan pola dapat digabungkan dengan metode prediksi seperti jaringan syaraf tiruan, *fuzzy*, dan *Support Vector Machine* (SVM). Tujuan dari penggabungan

[1]*Mahasiswa, Jurusan Teknik Elektro, Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember, Jl. Raya Sukolilo, Surabaya (tlp: 031-5947302; fax: 031-5931237; e-mail: berlian14@mhs.ee.its.ac.id)*

[2]*Dosen, Jurusan Teknik Elektro, Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember, Jl. Raya Sukolilo, Surabaya (tlp: 031-5947302; fax: 031-5931237; e-mail: sardjono@ee.its.ac.id, hery@ee.its.ac.id)*

[3]*Dosen, Jurusan Teknik Biomedik, Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember, Jl. Raya Sukolilo, Surabaya (tlp: 031-5947302; fax: 031-5931237; e-mail: sardjono@ee.its.ac.id)*

[4]*Dosen, Jurusan Teknik Komputer, Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember, Jl. Raya Sukolilo, Surabaya (tlp: 031-5947302; fax031-5931237; e-mail: hery@ee.its.ac.id)*

[5]*Dosen, Jurusan Teknik Informatika, Fakultas Teknik, Universitas 17 Agustus 1945 Surabaya, Jl. Raya Semolowaru 45, Surabaya (e-mail: berlian.alkindhi@untag-sby.ac.id)*

metode ini adalah untuk menganalisis pola dan mengelompokkan urutan nukleotida berdasarkan kesamaan dan keterkaitan tertentu. Pada makalah ini, digabungkan pengenalan pola dengan berbagai jenis metode SVM [9]. Tujuannya adalah memprediksi adanya infeksi HCV di dalam DNA manusia. Diuji coba enam jenis pendekatan SVM untuk menganalisis metode yang terbaik dari SVM untuk studi kasus penelitian ini.

Virus Hepatitis C adalah jenis virus RNA, yaitu virus yang bertanggung jawab pembawa pesan kode untuk pembentukan DNA/protein baru. Jika RNA terinfeksi virus, akibatnya DNA yang terbentuk juga akan berubah. Sifat alami sel adalah mempertahankan diri supaya tetap hidup atau malah mati dengan mengubah pola kode DNA yang baru melalui perubahan pola RNA. RNA yang terinfeksi HCV juga akan membentuk pola kode DNA yang berbeda dengan aslinya [10].

Pada mahluk hidup, secara alami dan otomatis, DNA abnormal seharusnya akan mati, tetapi sifat lainnya adalah berusaha beradaptasi dengan mengubah pola kode RNA sehingga DNA yang terbentuk dapat bertahan hidup. Sifat RNA yang dapat beradaptasi mengakibatkan pola kode DNA baru juga berubah. Siklus ini berjalan berantai terus menerus. DNA kode baru akan membuat sinyal RNA baru, RNA yang baru terinfeksi lagi dengan HCV akan membentuk DNA baru lagi, ke depannya RNA yang baru beradaptasi membentuk pola kode baru untuk DNA yang akan dibentuk dan berulang lagi. Dengan kondisi yang seperti ini, pada HCV pola DNA sel selalu dapat berubah-ubah (sebagian tetap dengan pola kode yang sama, sebagian mengalami perubahan). Pola-pola RNA inilah yang disebut subtipe genom dari HCV [11].

II. KLASIFIKASI *SEQUENCE* DNA

Salah satu tujuan pencarian suatu *sequence* di dalam *isolated DNA* adalah untuk menganalisis adanya suatu mutasi virus atau bakteri di dalam *isolated DNA*. Satu *isolated DNA* dapat terdiri atas puluhan ribu *sequence* nukleotida, dan dalam bank gen terdapat jutaan *isolated DNA*. Ini akan menyebabkan dibutuhkannya waktu yang lama jika pencarian *sequence* tersebut dilakukan secara manual. Selain itu, mutasi virus atau bakteri tersebut dapat berbeda-beda sesuai dengan cara RNA pada *isolated DNA* tersebut mempertahankan diri. Makalah ini mempelajari cara menentukan sebuah *sequence* tersebut adalah suatu mutasi atau bukan dan cara mengenalinya pada *isolated DNA*. Diterapkan metode SVM dengan menguji coba beberapa *kernel*-nya untuk mengetahui *kernel* SVM yang paling bagus kinerjanya untuk studi kasus ini.

Pada algoritme SVM, waktu yang diperlukan untuk mengklasifikasikan titik data yang tidak diketahui adalah proporsional dengan jumlah vektor dukungan [12], [13]. Tergantung pada kompleksitas struktur kelas, kadang-kadang jumlah vektor dukungan dari model SVM meningkat dengan jumlah titik data pelatihan [14]. Salah satu solusinya adalah dengan mengurangi jumlah vektor dukungan, tetapi mempertahankan lebih sedikit tingkat akurasi yang sama seperti SVM normal yang tidak menggunakan pengurangan vektor dukungan [15]. Sebuah SVM menemukan *hyperplane* yang memisahkan memaksimalkan margin pemisahan dan karenanya, lokasi *hyperplane* terutama tergantung pada satu set "titik batas" [16]. Algoritme SVM dapat diterapkan pada data pelatihan yang dikurangi untuk menghasilkan model klasifikasi. Model klasifikasi ini dilakukan dengan menilai kinerja dengan melonggarkan definisi titik batas [17]. Selain itu, optimasi SVM dapat juga dilakukan dengan memperluas algoritme ke ruang fitur menggunakan transformasi *kernel*. Dalam hal ini, sebuah vektor dukungan dihasilkan dalam ruang fitur menggunakan matriks *kernel* terkait [18].

Terdapat dua cara yang standar dalam mesin pembelajaran, yaitu *supervised* dan *unsupervised*, tetapi beberapa literatur telah menemukan algoritme yang mampu menyelesaikan masalah data tak terstruktur, yaitu dengan pengubahan algoritme menjadi *semi-supervised* (semiterbimbing) [19], [20]. Pembelajaran semi-terbimbing adalah salah satu paradigma pembelajaran yang paling menjanjikan dalam banyak aplikasi praktis dengan beberapa sampel berupa data tak terstruktur. Di antara model pembelajaran semacam itu, SVM adalah yang paling umum dan menonjol [21]. Namun, SVM *semi-supervised* yang khas tidak dapat memperkirakan distribusi sampel positif dan negatif dengan baik. Salah satu solusi dari masalah tersebut adalah dengan menyajikan kombinasi dari dua strategi multiklasifikasi untuk mengurangi waktu berjalan dan meningkatkan akurasi klasifikasi secara bersamaan. Metode tersebut disebut juga dengan *ensemble S3 SVM* [22]. Metode SVM dapat menangani masalah klasifikasi semiterbimbing bahkan dengan distribusi yang tidak diketahui atau data yang tidak seimbang.

Dalam bioinformatika, SVM sering digunakan untuk klasifikasi suatu citra maupun data kesehatan. Selain data tidak terstruktur, data multidimensi juga merupakan salah satu faktor dibutuhkannya inovasi pada algoritme yang sudah ada. Telah diusulkan sebuah aplikasi yang mampu mereduksi data multidimensi, yaitu *Minimax Concave Ridge Support Vector Machine* (MCR SVM) yang secara bersamaan melakukan klasifikasi dan pengurangan dimensi [23]. Pengklasifikasi SVM yang diusulkan ini menggabungkan keuntungan dari ketidaksempurnaan pengestimasi SVM dan kemampuan seleksi grup fitur SVM untuk mengatasi kerugian. Selain itu, juga diberikan pembenaran teoretis untuk fitur sparsial yang dipilih.

Selain sebagai analisis citra dan data kesehatan, SVM juga mampu menentukan efektivitas pengujian obat di rumah sakit. Tugas yang paling membutuhkan waktu dari obat-obatan adalah mendiagnosis dan memilih pengobatan. Secara tradisional, dokter telah memecahkan masalah ini, hanya mengandalkan intuisi dan pengalaman. SVM mampu memberikan analisis diagnosis dan saran obat dengan melakukan pengelompokan berdasarkan gejalanya. Parameter yang digunakan untuk pengujian SVM ini adalah deskripsi matematis dan perumusan masalah. Tujuan pemanfaatan SVM pada sistem pakar adalah membuat prasyarat untuk diagnosis pencegahan penyakit pasien. Dengan menggunakan SVM dan sistem pemantauan, spesialis dapat mendiagnosis dan mengembangkan perawatan yang optimal dengan benar [24].

Metode SVM dapat digabungkan dengan metode yang lain sebagai aktivitas sebelum klasifikasi. SVM dapat digabungkan dengan metode *Convolutional Neural Network* (CNN) untuk mendeteksi adanya leukemia di dalam tubuh manusia [25].

CNN berfungsi sebagai penentu prediksinya, kemudian untuk membagi antara kelompok yang terjangkit leukemia dan bukan dapat digunakan metode SVM. Pada makalah ini, digabungkan metode *Edit Lvenshtein Distance* dengan optimasi SVM. *Edit Levenshtein Distance* berfungsi untuk menganalisis adanya pola di dalam *isolated DNA*, kemudian hasil dari pengenalan pola tersebut akan menjadi matriks set data sebagai masukan dari SVM. Studi literatur pada makalah ini adalah sebagai gambaran penelitian tentang DNA yang menerapkan SVM dan melalui studi literatur tersebut, dilakukan penelitian ini untuk menganalisis kinerja terbaik dari *kernel* SVM untuk studi kasus ini.

III. METODOLOGI

Pada bagian ini dijelaskan keenam jenis metode SVM yang diujikan, yaitu *Linear SVM, Quadratic SVM, Cubic SVM, Fine Quadratic Gaussian SVM, Medium Gaussian SVM,* dan *Coarse Gaussian SVM*. Perbedaan dari keenam jenis metode SVM ini adalah cara menentukan *hyperplane*-nya. *Hyperplane* adalah kunci dalam SVM untuk menentukan dalam kelompok mana sebuah *node* tergabung.

Data set yang digunakan adalah 1.000 *isolated DNA* yang terdiri atas 500 *isolated HCV* dan 500 *isolated Homo Sapiens*. Dalam satu *isolated DNA*, terdapat 10.000 hingga 15.000 *sequence* DNA. Data tersebut diunduh dari bank gen dunia [26]. Kemudian sebagai parameter pembanding adalah primer HCV yang diperoleh dari Institut Penyakit Tropis, Universitas Airlangga. Setiap data *isolated DNA* dinormalisasi ke dalam bentuk FASTA terlebih dahulu. Kemudian, data FASTA tersebut dihitung nilai kemiripannya menggunakan metode *Edit Levenshtein Distance*. Hasil dari pengolahan metode *Edit Levenshtein Distance* tersebut dimasukkan ke dalam data matriks sebagai variabel *x* dalam SVM, sehingga besarnya matriks data tersebut adalah sebesar jumlah pola dikalikan dengan jumlah *isolated DNA*, yaitu 37 x 1.000. Sedangkan target atau variabel *y* adalah nilai positif atau negatifnya *isolated DNA* tersebut terhadap HCV dan besarnya matriks target adalah 1 x 1.000.

Data set pada makalah ini dapat diterapkan pada SVM sesuai dengan (1), dengan y_i adalah nilai antara -1 dan 1 yang mengindikasikan kelas dari x_i.

$$(x_i, y_1), .., (x_n, y_n)$$

$$\mathrm{x} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}; \; y = \begin{bmatrix} y_1 \\ .. \\ y_n \end{bmatrix}. \quad (1)$$

A. Linear SVM

Linear SVM adalah algoritme pembelajaran mesin untuk memecahkan masalah klasifikasi multikelas dari kumpulan data ultra besar yang mengimplementasikan versi asli dari algoritme bidang pemotongan untuk merancang SVM yang linear. *Linear SVM* adalah rutinitas linear *scalable* yang berarti bahwa ia menciptakan model SVM dalam waktu CPU yang berskala secara linear dengan ukuran kumpulan data pelatihan. Perbandingan dengan model SVM yang dikenal lainnya dengan jelas menunjukkan kinerja yang unggul ketika akurasi tinggi diperlukan.

Pada *linear SVM*, nilai x_i adalah *p-dimensional* dari vektor, sehingga untuk mendapatkan jarak terdekat *hyperplane* ke dua grup, yaitu antara x_i dengan nilai $y_i = 1$ dan $y_i = -1$ adalah dengan memaksimalkan margin dari *hyperplane.* Persamaan (2) adalah bentuk sederhana pemisahan *data set* x_i menjadi dua kelompok. Variabel *b* adalah konstanta yang menentukan lokasi pemisah terhadap titik asal.

$$\overrightarrow{w} \cdot \vec{x} \; - \; b \; = 0. \quad (2)$$

Ketika *data set* yang ada tidak tentu, dengan hasil yang diinginkan adalah dapat terbagi secara linear, maka dapat digunakan persamaan linear yang adaptif seperti pada (3), dengan y_i adalah target yang diinginkan, dalam hal ini adalah 1 untuk positif dan -1 untuk negatif. Sedangkan $(\vec{w} \cdot \overrightarrow{x_i} - b)$ adalah keluaran dari penghitungan SVM.

$$max(0{,}1 - y_i(\vec{w} \cdot \overrightarrow{x_i} - b)). \quad (3)$$

B. Quadratic SVM

Quadratic SVM adalah salah satu pendekatan pada SVM yang menyederhanakan fungsi *kernel* dengan melakukan fungsi kuadrat. Fungsi kuadrat ini dapat diterapkan pada *data set* yang saling terkait maupun *data set* jenis *time series* [27]. Penerapan fungsi kuadratik SVM dapat diamati pada (4), dengan *T* adalah fungsi kuadrat urutan data atau *time series.*

$$\overrightarrow{w^T} \cdot x + b > 0, untuk \; y = 1$$

atau

$$\overrightarrow{w^T} \cdot x + b < 0, untuk \; y = -1.$$

Maka *hyperplane:*

$$\overrightarrow{w^T} \cdot x + b = 0, (\vec{w}, b). \quad (4)$$

C. Cubic SVM dan Fine Quadratic Gaussian SVM

Pendekatan SVM dapat diperluas ke permukaan nonlinear dengan menggunakan trik *kernel*. Telah diusulkan penggolongan SVM nonlinear dengan menggunakan trik *kernel* [16]. Fungsi nonlinear dapat memindahkan ruang asli ke ruang dimensi yang lebih tinggi. Trik *kernel* tersebut dapat diterapkan pada dua jenis metode SVM, yaitu *Cubic* dan *Quadratic Gaussian* dengan *d* = 3 untuk *Cubic* dan *d* = 2 untuk *Quadratic Gaussian.* Trik *kernel* SVM untuk kedua metode dapat diamati pada (5).

$$k(x_i, x_j) = (x_i, x_j)^d. \quad (5)$$

D. Medium dan Coarse Gaussian SVM

Fungsi *Gaussian* pada SVM dikenal juga dengan fungsi radikal basis. *Gaussian SVM* termasuk salah satu jenis SVM nonlinear, yaitu setiap titiknya ditentukan oleh *non-linear kernel function.* Penghitungan trik *kernel* dilakukan berdasarkan fungsi Gauss yang diterapkan pada *hyperplane.* Yang membedakan antara *medium* dan *coarse Gaussian* adalah pada cara penghitungan *hyperplane.* Hal ini memungkinkan didapatkannya *hyperplane* yang cocok untuk semua anggota, dan hasil penempatan *hyperplane* akan

berbentuk kurva Gauss. Penghitungan Gaussian atau radikal basis dapat diamati pada (6).

$$k(x_i, x_j) = exp\ (-\gamma \|x_i, x_j\|^2). \quad (6)$$

Pada (6) dapat diterapkan untuk $\gamma = 0$, jika parameternya berbeda maka $\gamma = 1/\ (2\sigma^2)$.

E. Optimasi SVM

Dari enam metode yang diuji coba, ditambahkan fungsi optimasi agar hasil pengujian dapat sesuai harapan. Optimasi tersebut adalah dengan menghitung nilai *hyperplane* seminimal mungkin. Secara umum, jarak antara dua *hyperplane* adalah $2/\|\vec{w}\|$, sehingga untuk memaksimalkan jarak antara dua *hyperplane* adalah dengan meminimalkan nilai $\|\vec{w}\|$. Optimasi SVM dapat diamati pada (7), dengan γ adalah parameter yang menentukan antara meningkatkan ukuran margin dan memastikan bahwa $\vec{x_i}$ berada pada sisi yang benar dari margin.

$$y_i(\vec{w} \cdot \vec{x_i} - b) \geq 1, untuk\ i = 1, .., n$$

sehingga untuk menentukan *hyperplane* dapat menjadi:

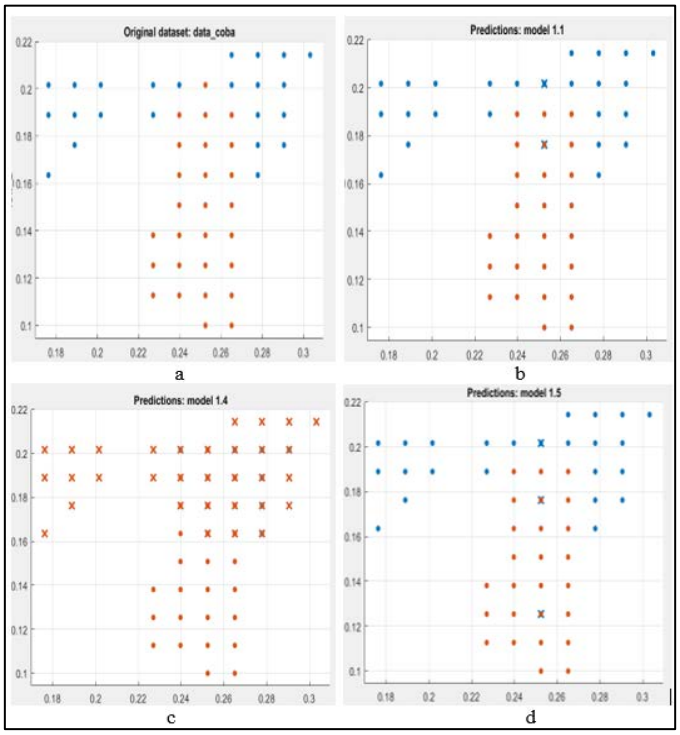
$$\left[\frac{1}{n} \sum_{i=1}^{n} max\ (0, 1 - y_i(\vec{w} \cdot \vec{x_i} - b))\right] + \gamma \|\vec{w}\|^2. \quad (7)$$

IV. ANALISIS HASIL

Seluruh metode SVM yang telah dijelaskan pada bagian metodologi diujikan menggunakan *data set* yang sebelumnya telah dinormalisasi ke dalam matriks berukuran 38 x 1.000. Variabel pelatihan sebanyak 37 dan variabel targetnya adalah data ke-38.

Data set tersebut berasal dari proses pengenalan pola pada *isolated DNA* menggunakan metode *Edit Levenshtein Distance*, kemudian pada masing-masing *isolated DNA*, dari seluruh *sequence* yang diujikan terhadap primer, diambil jarak terpendek terhadap masing-masing *isolated DNA*. Jarak terpendek tersebut yang menjadi nilai variabel *x* pada SVM, sedangkan variabel *y* adalah nilai prediksi *isolated DNA* tersebut positif atau negatif, yaitu 1 dan -1.

Hasil pemetaan data dapat diamati pada Gbr. 1, dengan plot biru adalah data negatif (grup 0) dan plot merah adalah positif (grup 1). Gbr. 1(a) menunjukkan plot data asli sebelum mengalami pemisahan kelompok dengan SVM, dengan plot merah dan plot biru terpisah dengan jelas. Gbr. 1(b) adalah pemetaan plot data dari keempat metode, yaitu *Linear*, *Quadratic*, *Cubic*, dan *Coarse Gaussian SVM*. Karena hasil pemisahan datanya sama, maka pemetaan plot datanya pun juga sama. Gbr. 1(b) memiliki dua plot *error*, yaitu data yang seharusnya bernilai positif tetapi mendapat hasil prediksi SVM bernilai negatif. Gbr. 1(c) adalah pemetaan plot data metode *Fine Gaussian SVM*. Pada gambar ini, tampak bahwa banyak sekali data yang mengalami kesalahan prediksi. Gbr. 1(d) adalah pemetaan plot data *Medium Gaussian SVM*. Tampak pada gambar tersebut, terdapat tiga plot *error*, yaitu kesalahan prediksi, yang seharusnya bernilai positif diprediksi menjadi negatif.



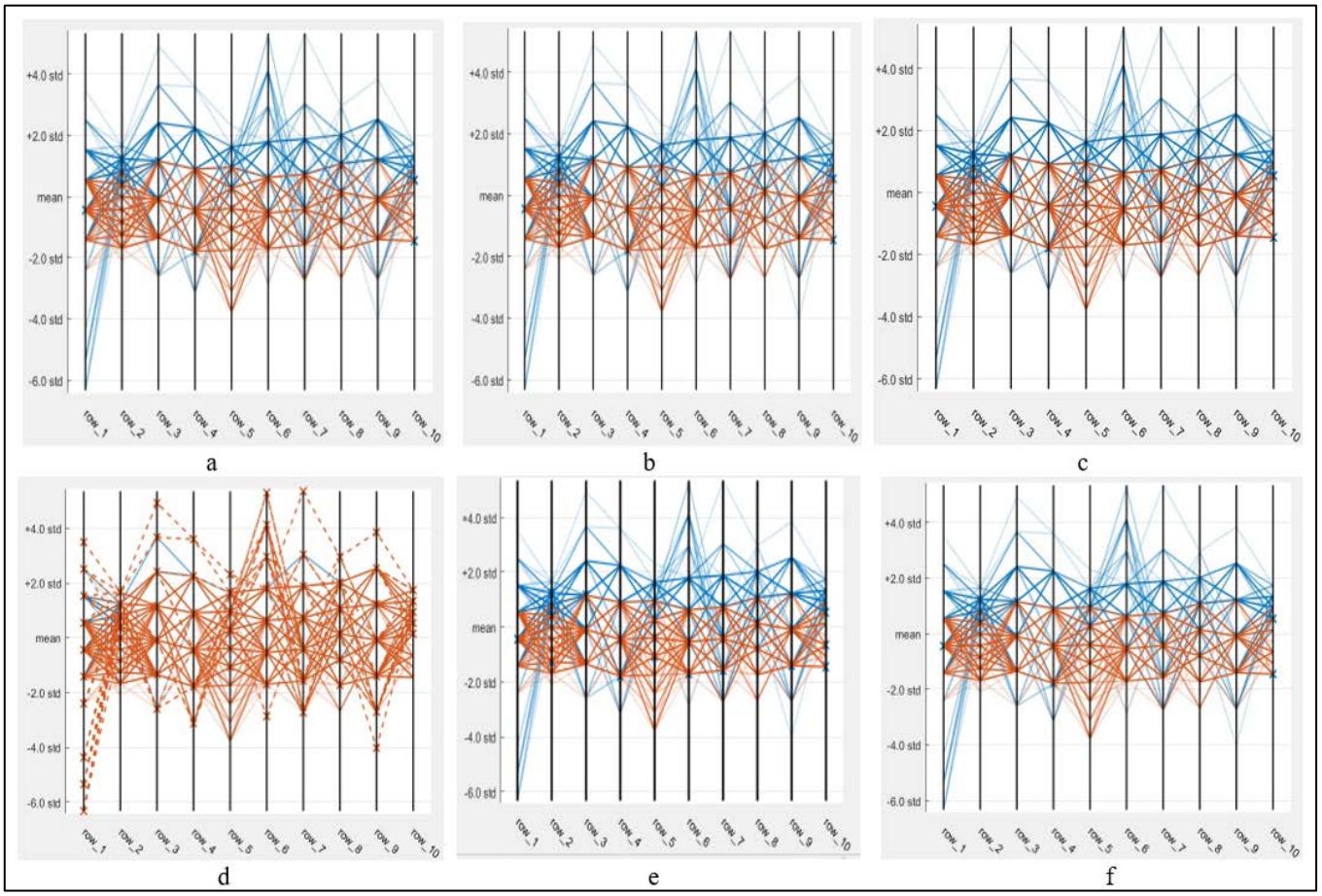
Gbr. 1 Pemetaan plot data pada SVM, (a) data asli, (b) plot data *Linear*, *Quadratic*, *Cubic*, *Coarse Gaussian*, (c) plot data *Fine Gaussian*, (d) plot data *Medium Gaussian*.

Untuk dapat lebih memperjelas kinerja masing-masing metode SVM, data disajikan pada Tabel I. Analisis yang dijelaskan pada Tabel I merupakan hasil pengolahan algoritme SVM menggunakan *data set*.

TABEL I
ANALISIS KECEPATAN PREDIKSI MASING-MASING METODE SVM

No.	Metode	Kecepatan Prediksi (obs/dtk)	Waktu Pelatihan (detik)
1	*Linear SVM*	~ 6.000	4,4208
2	*Quadratic SVM*	~ 19.000	0,84011
3	*Cubic SVM*	~19,000	0,74427
4	*Fine Gaussian SVM*	~ 15.000	1,2663
5	*Medium Gaussian SVM*	~17.000	0,71109
6	*Coarse Gaussian SVM*	~ 18.000	0,69129

Tabel I menunjukkan kecepatan masing-masing metode dalam mengolah *data set*. Dapat diamati bahwa metode *Linear SVM* memiliki kecepatan prediksi paling lambat, yaitu sekitar 6.000 obs/detik, dan waktu yang dibutuhkan untuk melatih data jauh lebih lama dibandingkan dengan metode yang lain. Waktu yang dibutuhkan metode *Linear SVM* untuk melakukan pelatihan *data set* adalah sebesar 4,4208 detik, atau setara dengan 3-7 kali lebih lambat dibandingkan dengan metode lainnya. Metode *Quadratic SVM* memiliki kecepatan prediksi paling tinggi, sedangkan waktu pelatihan paling rendah adalah metode *Fine Gaussian SVM*.

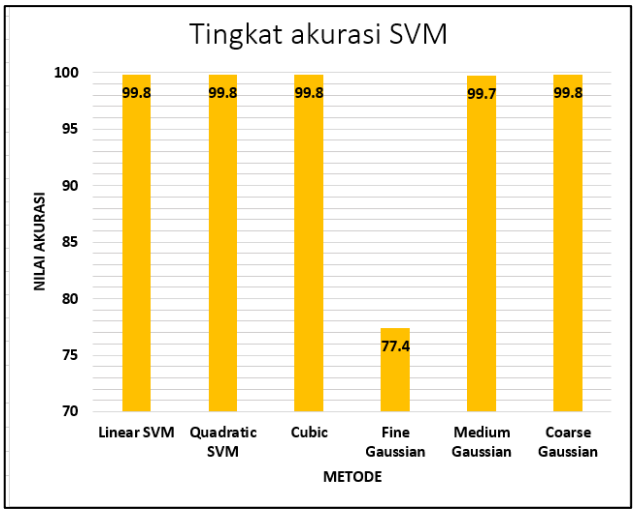


Gbr. 2 Grafik paralel koordinat plot masing-masing metode SVM, (a) *Linear SVM*, (b) *Quadratic SVM*, (c) *Cubic SVM*, (d) *Fine Gaussian SVM*, (e) *Medium Gaussian*, (f) *Coarse Gaussian SVM*.

Dari keenam metode yang diujikan, keseluruhannya mampu mengolah *data set*. Hasil pengujian dapat diamati pada Gbr. 2, dengan setiap plot merah dan biru dapat terpisah secara kontinu. Sedangkan untuk metode *Fine Gaussian SVM*, tampak bahwa plot merah dan plot biru saling bertumpukan (tidak terpisah) sehingga yang tampak adalah koordinat plot merah saja. Hal ini dikarenakan *Fine Gaussian SVM* tidak mampu memisahkan data dengan baik, sehingga nilai y = -1 dan y = 1 bertumpuk menjadi satu. Bentuk *kernel Fine Gaussian* untuk memisahkan dua data adalah seperti kurva lengkung yang berdasarkan fungsi Gauss.

V. PEMBAHASAN

Pengukuran akurasi dari hasil percobaan dalam makalah ini menggunakan metode *10 K-Fold Cross Validation*, yaitu 1.000 *data set* dibagi menjadi 10 kelompok secara acak. Tiap satu kelompok terdiri atas 100 data sebagai target dan 900 data sisanya adalah sebagai data pelatihan. Proses ini berulang sebanyak jumlah kelompok yang ditentukan hingga seluruh kelompok telah menjadi data target pada prose SVM. Tujuan dari pengujian ini adalah untuk mengukur tingkat akurasi dari masing-masing metode SVM yang telah dioptimasi. Hasil pengolahan metode SVM dapat diamati pada Gbr. 3, dengan empat dari keenam metode memiliki nilai akurasi sebesar 99,8%, satu di antaranya 99,7%, dan satu lainnya sebesar 77,4%, yaitu metode *Fine Gaussian SVM*. Nilai akurasi diperoleh dengan menjumlahkan seluruh data *True Positif* (TP) dan *True Negative* (TN), kemudian dibagi dengan jumlah seluruh data yang diuji. TP adalah jumlah data yang diprediksi positif HCV dan memang sesuai data target, data tersebut positif. Sebaliknya, TN adalah jumlah data yang diprediksi negatif (Homo Sapiens normal) dan memang sesuai data target, data tersebut negatif HCV.



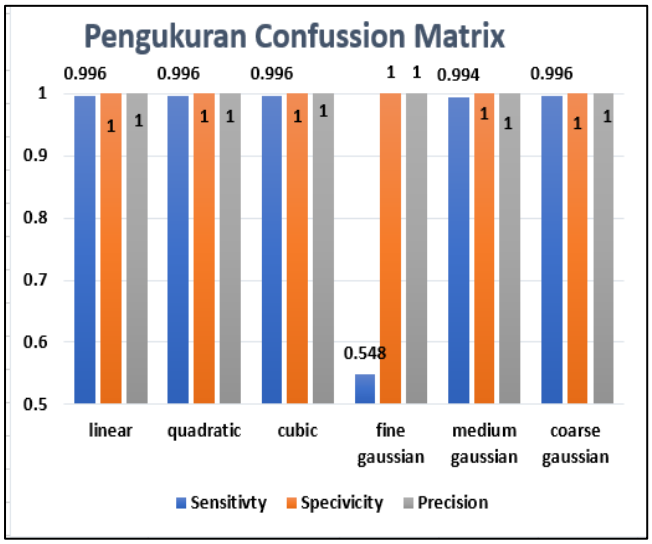
Gbr. 3 Tingkat akurasi masing-masing metode SVM.

HCV memiliki tingkat mutasi yang tinggi, sehingga terdapat banyak sekali subtipe pada HCV dan setiap subtipe dapat menghasilkan lebih dari satu primer. Banyaknya variasi primer akibat mutasi HCV menyebabkan proses pencarian pola *sequence* DNA menjadi sulit. Oleh karena itu, selain pemilihan *kernel* SVM yang terbaik, juga dibutuhkan *data set* yang banyak. Semakin banyak *data set* yang digunakan, proses pelatihan pada mesin pembelajaran akan semakin bagus, sehingga menghasilkan proses pengujian yang bagus juga. Pada makalah ini, *data set* mencapai 1 milyar *sequence* DNA yang terdiri atas *sequence* Homo Sapiens dan *sequence* HCV.

TABEL II
PENGUKURAN AKURASI MASING-MASING METODE SVM

No.	Metode	TP	TN	FP	FN
1	*Linear SVM*	498	500	2	0
2	*Quadratic SVM*	498	500	2	0
3	*Cubic SVM*	498	500	2	0
4	*Fine Gaussian SVM*	274	500	226	0
5	*Medium Gaussian SVM*	497	500	3	0
6	*Coarse Gaussian SVM*	498	500	2	0

Analisis jumlah *data set* dengan prediksi benar dan salah dapat diamati pada Tabel II. TP adalah *True Positive* yaitu data yang dianggap positif oleh SVM dan memang data tersebut positif. TN adalah *True Negative*, yaitu data yang dianggap negatif oleh hasil pengolah SVM dan memang data tersebut negatif. FP adalah *False Positive* yaitu data yang dianggap positif oleh SVM padahal data tersebut sebenarnya negatif. FN adalah *False Negative*, yaitu data yang dianggap negatif oleh SVM padahal data tersebut sebenarnya positif. Melalui hasil TP, TN, FP, dan FN, dapat dihitung nilai sensitivitas, spesifik, dan presisi. Hasil penghitungan sensitivitas tersebut dapat diamati dalam bentuk grafik pada Gbr. 4.



Gbr. 4 *Bar* sensitivitas (warna biru), spesifik (warna merah), dan presisi (warna abu-abu) masing-masing metode.

Nilai sensitivitas digunakan untuk menganalisis besar data yang terdeteksi positif dengan benar dibandingkan data yang teridentifikasi negatif tetapi ternyata pada target sebenarnya positif. Sebaliknya, nilai spesifik digunakan untuk menganalisis besar data yang terdeteksi negatif dengan benar dibandingkan data yang teridentifikasi positif tetapi ternyata pada target sebenarnya negatif. Untuk melengkapi kedua *confusion matrix* tersebut, ditambahkan analisis presisi, yaitu analisis jumlah data positif dibandingkan dengan jumlah seluruh hasil prediksi yang bernilai positif.

VI. KESIMPULAN

Metode SVM banyak diterapkan untuk memisahkan set data yang berdasarkan *hyperplane*, yaitu garis yang membagi dua buah kelompok. Untuk memisahkan set data, SVM memiliki beberapa pendekatan berdasarkan cara *hyperplane* membagi set data menjadi dua buah kelompok. Pada makalah ini, digunakan beberapa pendekatan SVM, antara lain *Linear SVM*, *Quadratic SVM*, *Cubic SVM*, *Fine Gaussian SVM*, *Medium Gaussian SVM*, dan *Coarse Gaussian SVM*. Keenam jenis pendekatan tersebut diuji untuk mendapatkan model *hyperplane* yang terbaik untuk memisahkan set data DNA. Set data berupa 1.000 *isolated DNA*, dan dibangun mesin pembelajaran yang mampu mengenali *isolated DNA* yang positif terinfeksi HCV,

Makalah ini bertujuan untuk mencari kinerja paling baik antara keenam *kernel* SVM jika diterapkan pada pengelompokan DNA. Hasil pengujian menunjukkan, semua metode menghasilkan nilai kinerja yang tinggi, yaitu 99,8%. Dengan kata lain, dari seribu data hanya terdapat dua data yang *error* prediksi. Metode *Fine Gaussian SVM* adalah metode SVM yang memiliki kinerja paling rendah pada studi kasus ini, yaitu sebesar 77,4% dengan 226 data *error* prediksi. Uji coba ini membuktikan bahwa sebagian besar pendekatan SVM mampu mengenali adanya mutasi HCV di dalam *isolated DNA*. Rencana penelitian lanjutan adalah akan diuji coba berbagai metode prediksi dan pengelompokan yang hasilnya akan dibandingkan dengan SVM untuk membuktikan metode yang terbaik untuk studi kasus penelitian ini.

REFERENSI

[1] Jun Hu, Yang Li, Ming Zhang, Xibei Yang, Hong-Bin Shen, dan Dong-Jun Yu, ”Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-Based Features and Boosting Multiple SVMs,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 14, No. 6, hal. 1389-1398, 2017.

[2] Bin Liu, Shanyi Wang, Qiwen Dong, Shumin Li, Xuan Liu, ”Identification of DNA-Binding Proteins by Combining Auto-Cross Covariance Transformation and Ensemble Learning,” *IEEE Transactions on NanoBioscience*, Vol. 15, No. 4, hal. 328-334, 2016.

[3] Jianmin Ma, Minh N. Nguyen, dan Jagath C. Rajapakse, ”Gene Classification Using COdon USage and Support Vector Machine,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 6, No. 1, hal. 134-143, 2009.

[4] Lakshmi Chaitanya, Krystal Breslin, Sofia Zuñiga, Laura Wirken, Ewelina Pośpiech, Magdalena Kukla-Bartoszek, Titia Sijen, Peter de Knijff, Fan Liu, Wojciech Branicki, Manfred Kayser, Susan Walsh, ”The HIrisPlex-S System for Eye, Hair and Skin Colour Prediction from DNA: Introduction and Forensic Developmental Validation,” *Forensic Science International: Genetics*, Vol. 35, hal. 123-135, 2018.

[5] Javier R. Revollo, Azra Dad, Lea P. McDaniel, Mason G. Pearce, Vasily N. Dobrovolsky, ”Genome-Wide Mutation Detection by Interclonal Genetic Variation,” *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, Vol. 829-830, hal. 61-69, 2018.

[6] Luis Alberto dan Hernandez Montiel, ”Hybrid Algorithm Applied on Gene Selection and Classification from Different Siseases,” *IEEE Latin America Transactions*, Vol. 14, No. 2, hal. 930-935, 2016.

[7] Berlian Al Kindhi, Tri Arief Sardjono, ”Pattern Matching Performance Comparisons as Big Data Analysis Recommendation for Hepatitis C Virus (HCV) Sequence DNA,” *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, Kinabalu, 2015, hal. 99-104.

[8] Berlian Al Kindhi, Muhammad Afif Hendrawan, Diana Purwitasari, Tri Arief Sardjono, Mauridhi Hery Purnomo, ”Distance-Based Pattern Matching of DNA Sequences for Evaluating Primary Mutation,” *2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, 2017, hal. 310-314.

[9] Annisa Handayani, Ade Jamal, Ali Akbar Septiandri, ”Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara,” *JNTETI (Jurnal Nasional Teknik Elektro dan Teknologi Informasi)*, Vol. 6, No. 4, hal. 394-403, 2017.

[10] Maria Pujantell, Sandra Franco, Iván Galván-Femení, Roger Badia, Marc Castellví, Edurne Garcia-Vidal, Bonaventura Cloteta, Rafael d Cid, Cristina Tural, Miguel A.Martínez, Eva Riveira-Muñoz, José A.Esté, Ester Ballana, ”ADAR1 Affects HCV Infection by Modulating Innate Immune Response,” *Antiviral Research*, Vol. 156, hal. 116-127, 2018.

[11] Annettevon Delft, Timothy A.Donnison, José Lourenço, Claire Hutchings, Caitlin E.Mullarkey, Anthony Brown, Oliver G.Pybus, Paul Klenerman, Senthil Chinnakannan, Eleanor Barnes, ”The Generation of a Simian Adenoviral Vectored HCV Vaccine Encoding Genetically Conserved Gene Segments to Target Multiple HCV Genotypes,” *Vaccine*, Vol. 36, No. 2, hal. 313-321, 2018.

[12] Sebastián Maldonado, Julio López, ”Dealing with High-Dimensional Class-Imbalanced Datasets: Embedded Feature Selection for SVM Classification,” *Applied Soft Computing*, Vol. 67, hal. 94-105, 2018.

[13] David de la Mata-Moya, María Pilar Jarabo-Amores, Jaime Martín de Nicolás, Manuel Rosa-Zurera, ”Approximating the Neyman–Pearson Detector with 2C-SVMs. Application to Radar Detection,” *Signal Processing*, Vol. 131, hal. 364-375, 2017.

[14] Deepak Kumar Jain, Surendra Bilouhan Dubey, Rishin Kumar Choubey, Amit Sinhal, Siddharth Kumar Arjari, Amar Jain, Haoxiang Wang, ”An Approach for Hyperspectral Image Classification by Optimizing SVM Using Self Organizing Map,” *Journal of Computational Science*, Vol. 25, hal. 252-259, 2018.

[15] Rupan Panja, Nikhil R. Pal, ”MS-SVM: Minimally Spanned Support Vector Machine,” *Applied Soft Computing*, Vol. 64, hal. 356-365, 2018.

[16] Saurabh Paul, Malik Magdon-Ismail, Petros Drineas, ”Feature Selection for Linear SVM with Provable Guarantees,” *Pattern Recognition*, Vol. 60, hal. 205-214, 2016.

[17] M. A. Ebrahimi, M. H. Khoshtaghaza, S. Minaei, B. Jamshidi, ”Vision-based Pest Detection Based on SVM Classification Method,” *Computers and Electronics in Agriculture*, Vol. 137, hal. 52-58, 2017.

[18] Samia Djemai, Belkacem Brahmi, Mohand Ouamer Bibi, ”A Primal–dual Method for SVM Training,” *Neurocomputing*, Vol. 211, hal. 34-40, 2016.

[19] Yong Liu, Shizhong Liao, ”Granularity Selection for Cross-Validation of SVM,” *Information Sciences*, Vol. 378, hal. 475-483, 2017.

[20] Sidheswar Routray, Arun Kumar Ray, Chandrabhanu Mishra, G. Palai, ”Efficient Hybrid Image Denoising Scheme Based on SVM Classification,” *Optik*, Vol. 157, hal. 503-511, 2018.

[21] Jing Zhou, Ying Yang, Steven X. Ding, Yanyang Zi, Muheng Wei, ”A Fault Detection and Health Monitoring Scheme for Ship Propulsion Systems Using SVM Technique,” *IEEE Access*, Vol. 6, hal. 16207-16215, 2018.

[22] Dan Zhang, Licheng Jiao, Xue Bai, Shuang Wang, Biao Hou, ”A Robust Semi-Supervised SVM Via Ensemble Learning,” *Applied Soft Computing*, Vol. 65, hal. 632-643, 2018.

[23] Jian-wei Liu, Li-peng Cui, Xiong-lin Luo, ”MCR SVM Classifier with Group Sparsity,” *Optik - International Journal for Light and Electron Optics*, Vol. 127, No. 17, hal. 6915-6926, 2016.

[24] C. Venkatesan, P. Karthigaikumar, Anand Paul, S. Satheeskumaran, R. Kumar, ”ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications,” *IEEE Access*, Vol. 6, hal. 9767-9773, 2018.

[25] Luis H.S.Vogado, Rodrigo M.S. Veras, Flavio. H.D. Araujo, Romuere R.V. Silva, Kelson R.T. Aires, ”Leukemia Diagnosis in Blood Slides Using Transfer Learning in Cnns and SVM for Classification,” *Engineering Applications of Artificial Intelligence*, Vol. 72, hal. 415-422, 2018.

[26] (2018) NCBI website. [Online], ”www.ncbi.nlh.gov”, tanggal akses: 27 Mei 2018.

[27] I. Dagher, ”Quadratic Kernel-Free Non-Linear Support Vector Machine,” *Journal of Global Optimization*, Vol. 41, No. 1, hal. 15–30, 2008.