JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume XIII No 1, Juni 2024
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

IDEKTIFIKASI GEN MARKER PBMCS ISCHEMIC STROKE MENGGUNAKAN ANALISIS BIOINFORMATIKA DAN SUPPORT VECTOR MACHINE

M. Fauzan Azhari[1)], Rohmatul Fajriyah[2)]
Fakultas Teknologi Industri[(1,2)]
Fakultas Matematika dan Ilmu Pengetahuan Alam[(2)]
Universitas Islam Indonesia, Jl Kaliurang Km. 14,5 Sleman, Daerah Istimewa Yogyakarta
email: 21917011@students.uii.ac.id[1)], rfajriyah@uii.ac.id [2)]

Abstrak

Penyakit stroke adalah kondisi ketika aliran darah ke otak terhambat atau terputus, mengakibatkan kerusakan pada sel-sel otak. Diperkirakan ada 50 juta kasus stroke di seluruh dunia, dengan 9 juta di antaranya mengakibatkan kecacatan berat. Penelitian ini bertujuan untuk melakukan klasifikasi dan melihat genomic profiling antara penderita stroke iskemik kontrol dan non kontrol dengan analisis support vector machine (SVM). Data yang digunakan adalah data *microarray* dengan kode series GSE22255 dari Institut Kedokteran Molekuler di kota Lisbon, Portugal. Untuk melihat perbandingan akurasi yang dihasilkan, analisis dilakukan dengan beberapa skema berdasarkan kernel dan nilai *cost optimal* pada metode SVM yaitu kernel linier, polinomial, RBF dna sigmoid. Dari hasil analisis diketahui bahwa metode SVM dengan skema kernel terbaik yaitu menggunakan kernel RBF dan *optimal cost* 1 dengan nilai akurasi sebesar 88.0%, Model SVM dengan kernel sigmoid tidak dapat digunakan untuk klasifikasi karena nilai akurasinya yang sangat rendah. Sementara itu, SVM dengan kernel linear dan polynomial masih tetap dapat digunakan karena nilai akurasinya >70% .

Kata Kunci: *bioinformatika, ekspresi gen, klasifikasi, stroke iskemik, support vector machine*.

1. Pendahuluan

Penyakit stroke adalah kondisi ketika aliran darah ke otak terhambat atau terputus, mengakibatkan kerusakan pada sel-sel otak [1]. Terdapat dua jenis utama stroke salah satunya yaitu iskemik. Stroke iskemik terjadi karena pembuluh darah yang memasok darah ke otak tersumbat. Stroke merupakan penyebab kematian ketiga tertinggi di dunia. Di rumah sakit, stroke menjadi penyebab kematian kedua setelah penyakit jantung koroner. Diperkirakan ada 50 juta kasus stroke di seluruh dunia, dengan 9 juta di antaranya mengakibatkan kecacatan berat. Selain itu, stroke adalah penyebab utama kecacatan jangka panjang dan meningkatkan risiko gangguan kognitif lebih tinggi dibandingkan dengan mereka yang tidak pernah mengalami stroke [2].

Stroke dapat terjadi karena berbagai faktor risiko. Beberapa di antaranya termasuk tekanan darah tinggi, kadar kolesterol tinggi, diabetes, obesitas, dan kebiasaan merokok. Gaya hidup yang kurang aktif, konsumsi alkohol berlebihan, serta pola makan yang buruk juga berperan dalam meningkatkan risiko stroke [3]. Terkait dengan gejalanya, stroke memiliki gejala bervariasi, tetapi tanda-tanda umumnya meliputi kesulitan berbicara atau memahami ucapan, mati rasa atau kelemahan pada wajah, lengan, atau kaki, serta gangguan berjalan. Penting untuk segera mencari pertolongan medis jika terdapat gejala stroke, karena penanganan cepat dapat meningkatkan peluang pemulihan dan mengurangi risiko kerusakan otak.

Salah satu kontribusi dunia ilmu komputer adalah melakukan klasifikasi data ekspresi gen penderita stroke iskemik untuk melihat genomic profiling menggunakan analisis bioinformatika. Bioinformatics atau bioinformatika merupakan disiplin ilmu yang membahas kebutuhan untuk mengelola dan menafsirkan data yang dalam dekade terakhir ini secara besar-besaran dihasilkan oleh penelitian genom. Disiplin ilmu ini mewakili konvergensi genomik, bioteknologi, teknologi informasi yang meliputi analisis dan interpretasi data, pemodelan fenomena biologis serta pengembangan algoritma dan statistik [4].

Berdasarkan latar belakang yang sudah dijelaskan maka akan dilakukan analisis klasifikasi dengan penerapan teknik bioinformatika. Klasifikasi ini bertujuan untuk menentukan apakah status seorang penderita stroke iskemik telah dikontrol atau tanpa dikontrol berdasarkan ekspresi gen. Analisis klasifikasi dilakukan dengan metode support vector machine (SVM), yang termasuk dalam supervised learning dan digunakan untuk keperluan klasifikasi serta regresi.

2. Landasan Teori

Bioinformatika

Bioinformatika adalah bidang yang mengintegrasikan biologi, informatika, dan teknologi informasi untuk menganalisis serta menginterpretasikan data biologis. Bidang ini terutama berfokus pada pemrosesan dan analisis data sekuens DNA, RNA, dan protein dengan menggunakan metode komputasional. Analisis bioinformatika

memfasilitasi identifikasi gen, penyusunan profil ekspresi gen, serta penemuan hubungan antar biomolekul. Dengan memanfaatkan algoritma dan perangkat lunak khusus, bioinformatika mendukung penelitian di biologi molekuler, genomik, dan proteomik, serta membantu pengembangan obat dan pemahaman penyakit [5] [6].

Akses yang mudah dan efisien terhadap sumber-sumber bioinformatika sangat krusial untuk penelitian di bidang yang membutuhkan studi yang sistematis. Tujuan utamanya adalah memberikan perspektif baru yang mendukung perkembangan bioteknologi di masa depan [7]. Informasi dalam bioinformatika didapatkan melalui analisis data biologis menggunakan komputer. Data ini meliputi kode genetik, hasil eksperimen dari berbagai sumber, literatur ilmiah dan statistik pasien [8].

Support Vector Machine

Support Vector Machine (SVM) adalah metode dalam pembelajaran mesin yang digunakan untuk tugas-tugas klasifikasi dan regresi. Teknik ini berfungsi dengan menemukan *hyperplane* optimal yang memisahkan data ke dalam kelas-kelas yang berbeda dengan margin maksimal. SVM mampu bekerja dengan baik pada data berdimensi tinggi namun juga efektif untuk dataset berukuran kecil hingga menengah. Dengan menggunakan trik kernel, SVM dapat menangani kasus *non-linear* dengan memetakan data ke ruang dimensi yang lebih tinggi, sehingga memungkinkan pemisahan data yang lebih efektif. Keunggulan utama SVM adalah kemampuannya dalam meminimalkan kesalahan klasifikasi dan menghindari *overfitting* [9] [10].

Optimasi pada setiap kernel dan nilai $Cost\ (C)$ dilakukan dengan *trial and error* [11]. Berikut ini adalah beberapa kernel yang umum digunakan pada SVM.

a. *Polynomial*
Kernel ini cocok digunakan untuk menyelesaikan masalah klasifikasi dimana kumpulan data latih sudah dalam keadaan normal. Parameter yang dapat dioptimalkan yaitu $Cost\ (C)$ dan $degree\ (p)$ [12]. Fungsi kernel *polynomial* derajat p dituliskan dalam persamaan berikut.

$$K(x_i x_j) = (x_i \cdot x_j + 1)^p \tag{2.1}$$

Dimana:
x_i = variabel data *i*
x_j = variabel data *j*
p = konstanta dengan nilai lenih dari 0

b. *Raidal Basis Function* (RBF) atau *Gaussian*
Kernel RBF atau *Gaussian* merupakan pilihan untuk memodelkan persoalan klasifikasi ketika data terpisah secara *non linear*. Parameter yang terdapat pada kernel RBF yaitu $Cost\ (C)$ dan $Gamma\ (\Upsilon)$ [11]. Fungsi pada kernel ini dituliskan pada persamaan berikut.

$$K(x_i \cdot x_j) = \exp\left(-\Upsilon \left\| x_i - x_j \right\|^2\right) \tag{2.2}$$

Dimana:
x_i = variabel data *i*
x_j = variabel data *j*
Υ = nilai gamma

c. *Linear*
Kernel *linear* merupakan kernel yang paling umum digunakan untuk tugas-tugas yang memiliki pola data linier positif maupun negatif. Kernel *linear* mempunyai parameter $Cost\ (C)$ [11]. Kernel *linear* ditulis dengan persamaan berikut.

$$K(x_i x_j) = x_i \cdot x_j \tag{2.3}$$

Dimana:
x_i = variabel data *i*
x_j = variabel data *j*

d. *Sigmoid*
Sama seperti kernel RBF, parameter yang digunakan pada kernel *sigmoid* yaitu nilai $Cost\ (C)$ dan $Gamma\ (\Upsilon)$. Untuk persamaan pada kernel *sigmoid* dapat dituliskan pada persemaan berikut.

$$K(x_i \cdot x_j) = \tanh\left(\Upsilon . x_i - x_j + r\right) \tag{2.4}$$

Dimana:
x_i = variabel data *i*
x_j = variabel data *j*
Υ = nilai gamma

Confusion Matrix

Terdapat beberapa metrik yang bisa dipakai untuk mengevaluasi performa suatu model klasifikasi, diantaranya yaitu *accuracy*, *precision* dan *recall*. [13]. *Accuracy* merupakan ukuran ketepatan antara data prediksi dan data aktual. Sedangkan *precision* merupakan ukuran ketepatan dalam klasifikasi. Untuk metrik *recall* menjelaskan ukuran yang digunakan untuk mengukur proporsi kelas positif aktual yang diidentifikasi dengan benar [14]. Metrik-metrik tersebut dapat dihitung menggunakan tabel *confusion matrix* [15]. Jika dalam sebuah model klasifikasi hanya terdapat dua kelas, maka confusion matrix dapat diilustrasikan seperti yang ditunjukkan pada Tabel 1 berikut.

Tabel 1. *Confusion matrix* pada model dengan 2 kelas

Data Aktual	Data Prediksi	
	Positive	**Negative**
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

Pada tabel 1 merupakan ilustrasi *confusion matrix* apabila suatu model klasifikasi hanya memiliki 2 kelas. Tabel tersebut berguna untuk menilai performa model klasifikasi dengan membandingkan prediksi model dengan data aktual. Terdapat empat kemungkinan hasil prediksi pada *confusion matrix* dengan 2 kelas, antara lain yaitu sebagai berikut:

a. TP (True Positive), merujuk pada jumlah data yang diklasifikasikan dengan tepat sebagai kelas positif oleh model.
b. FP (False Positive), merujuk pada jumlah data aktual yang termasuk dalam kelas negatif, tetapi keliru diklasifikasikan sebagai kelas positif oleh model..
c. TN (True Negative), merujuk pada jumlah data yang diklasifikasikan dengan tepat sebagai kelas negatif oleh model.
d. FN (False Negative), merujuk pada jumlah data aktual yang termasuk dalam kelas positif, tetapi keliru diklasifikasikan sebagai kelas negatif oleh model.

Berikutnya, untuk mengevaluasi kinerja model klasifikasi dengan menggunakan metrik-metrik seperti *accyracy*, *precision*, *recall*, *spesificity*, dan *F1 score* dapat dihitung menggunakan persamaan berikut.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Data} \times 100\% \quad (2.4)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\% \quad (2.5)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100\% \quad (2.6)$$

$$Spesificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \times 100\% \quad (2.7)$$

$$F1\ Score = \frac{2 \times (Precision \times recall)}{Precision + recall} \quad (2.8)$$

3. Metode Penelitian

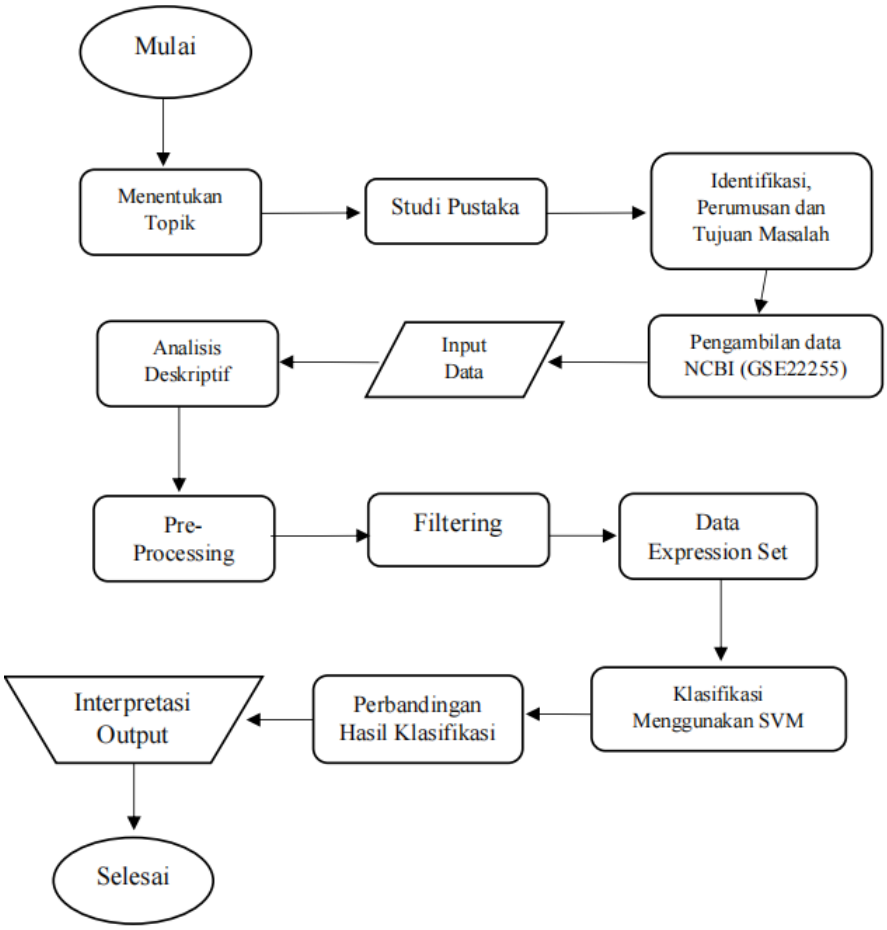
Sebelum memulai penelitian, langkah awal yang dilakukan oleh peneliti adalah menetapkan topik penelitian. Setelah melakukan riset dan berdasarkan latar belakang yang sudah dijelaskan sebelumnya, peneliti akan melakukan analisis bioinformatika dengan pendekatan machine learning terhadap data ekspresi gen. Selanjutnya peneliti melakukan pencarian studi pustaka atau referensi terkait yang relevan dengan topik tersebut. Dari topik penelitian dan studi pustaka yang telah dilakukan, maka peneliti dapat merumuskan tujuan penelitian yang ingin dicapai dan mencari data yang relevan dengan topik penelitian. Pada studi ini peneliti menggunakan pendekatan kuantitatif, dengan subjek dari penelitian ini adalah data ekspresi gen PBMCs stroke iskemik GSE22255 yang diperoleh melalui situs web www.ncbi.nlm.nih.gov/.

Setelah memperoleh data yang relevan dengan topik penelitian, langkah selanjutnya adalah menginput data tersebut ke dalam perangkat lunak yang digunakan (Rstudio) untuk analisis. Analisis awal yang dilakukan adalah analisis deskriptif untuk merangkum karakteristik dasar dari data yang akan digunakan. Pada tahap ini, peneliti menggunakan paket GEOquery dan affy untuk mengeksplorasi informasi yang tersedia dalam data tersebut.

Setelah melakukan analisis deskriptif, langkah selanjutnya yaitu melakukan pre-processing dan filtering. Pre-processing dan filtering dilakukan untuk menentukan variabel dalam data serta menyeleksi fitur yang akan digunakan untuk klasifikasi, termasuk menghitung standar deviasi dan nilai rata-rata. [16].

Selanjutnya, data yang telah melalui pre-processing dan filtering akan dimasukkan ke dalam set data yang baru, sehingga terbentuk data baru dalam bentuk data frame yang siap untuk dianalisis. Data tersebut selanjutnya dipisah menjadi dua bagian, yakni data untuk pelatihan (data latih) dan data untuk pengujian (data uji), dengan pembagian 80% untuk data latih dan 20% untuk data uji.

Langkah berikutnya adalah melakukan analisis pada data latih menggunakan metode klasifikasi SVM dengan berbagai kernel yang tersedia, selain itu pada penelitian ini penulis juga melakukan studi simulasi atau perbandingan terhadap nilai cost optimal pada masing-masing kernel. Model yang dibuat dari data latih akan dimanfaatkan untuk melakukan prediksi terhadap data uji, sehingga akan menghasilkan nilai akurasi prediksi dari masing-masing kernel. Setelah mendapatkan nilai akurasi dari setiap kernel, langkah berikutnya adalah membandingkan kernel mana yang menunjukkan akurasi terbaik dalam mengklasifikasikan data uji dan kemudian melakukan interpretasi berdasarkan model dengan kernel terbaik yang telah diidentifikasi. Detail langkah-langkah penelitian disajikan secara lebih rinci dalam diagram alir pada gambar 1.



Gambar 1. *Flowchart* penelitian

4. Hasil Penelitian

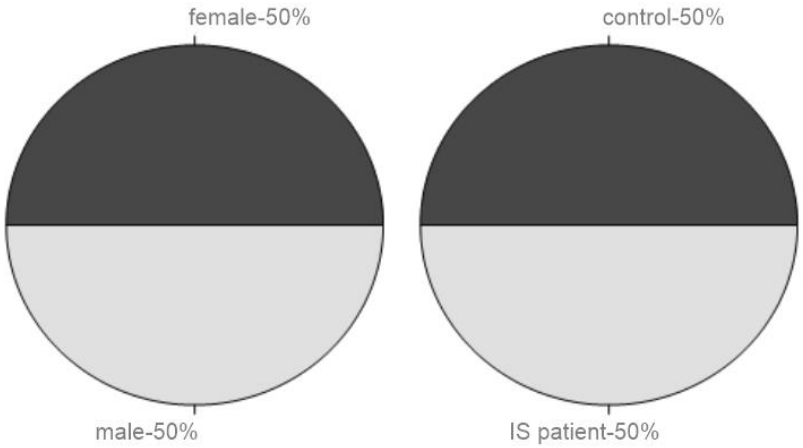
Analisis Deskriptif

Analisis deskriptif merupakan teknik dalam menganalisis data yang bertujuan untuk menggambarkan atau merangkum karakteristik dasar dari kumpulan data. Tujuan utama analisis deskriptif adalah memberikan gambaran yang jelas dan ringkas mengenai data tanpa melakukan inferensi atau kesimpulan yang lebih lanjut. Metode ini sering digunakan sebagai langkah awal dalam proses analisis data dan membantu peneliti untuk memahami pola, tren, dan distribusi data yang ada.

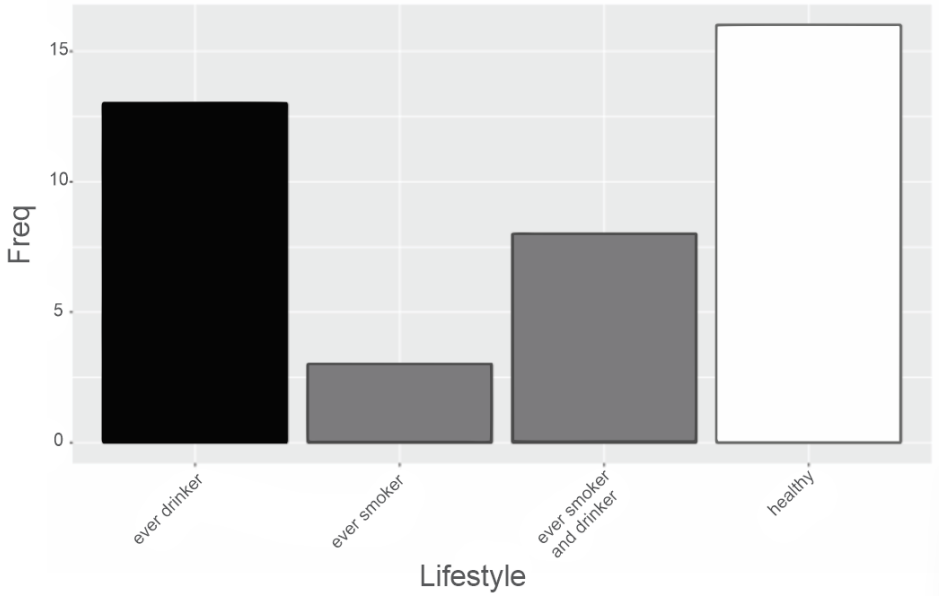
Pada data series GSE22255 terdapat sebanyak 40 sampel. Berdasarkan analisis deskriptif diketahui bahwa pasien laki-laki dan perempuan sama banyaknya yaitu berjumlah 20 orang. Demikian juga dengan status pasien, jumlah pasien dengan penyakit stroke iskemik sama dengan pasien tanpa penyakit stroke iskemik atau pasien kontrol yaitu 20 orang.

Kemudian banyaknya pasien yang mempunyai gaya hidup tidak sehat (*unhealthy lifestyle*) seperti merokok dan minum alkohol yaitu sebanyak 24 orang dan pasien dengan gaya hidup sehat (*healthy lifestyle*) sebanyak 16 orang. Selanjutnya berdasarkan keadaan klinis pasien, banyaknya pasien yang mempunyai penyakit penyerta seperti diabetes, *hypercholestrolemia* dan hipertensi berjumlah 27 orang dan pasien tanpa penyakit penyerta sebanyak 13 orang. Selain itu juga diketahui bahwa rentang usia pasien antara 45-74 tahun karena penyakit stroke

iskemik merupakan penyakit yang memiliki resiko tinggi bagi orang dengan usia lebih dari 40 tahun. Secara visual, karakteristik pasien ditampilkan pada gambar 1 dan gambar 2.



Gambar 2. Karakteristik data pasien berdasarkan gender dan status

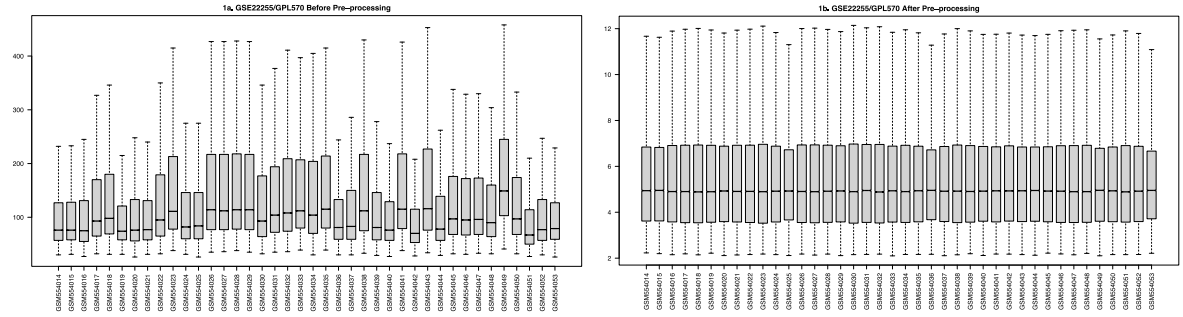


Gambar 3. Karakteristik data pasien berdasarkan pola hidup

Pre-processing

Hasil dari tahap *pre-processing* ditampilkan dalam bentuk *boxplot* pada gambar 4. *Bolxplot* merupakan salah satu jenis grafik yang digunakan dalam statistik untuk menggambarkan distribusi data numerik melalui kuartil. *Boxplot* memberikan visualisasi ringkas dari lima poin utama dalam data yaitu nilai minimum, kuartil pertama (Q1), median (Q2), kuartil ketiga (Q3) dan nilai maksimum. Secara visual, *boxplot* terdiri dari beberapa komponen seperti *box*, garis tengah (*median line*), garis antena (*whiskers*) dan *outliers*. *Box* menampilkan rentang antar kuartil atau *interquartile range* (*IQR*), yaitu rentang data data antara Q1 dan Q3 dan mencakup 50% data tengah. Garis tengah (*median line*) merupakan garis di dalam kotak yang menunjukan median atau nilai tengah dari data (Q2). Garis antena (*whiskers*) merupakan garus yang memanjang dari kedua ujung kotak yang menghubungkan antara nilai minimum dan maksimum yang bukan termasuk outlier. Sedangkan *outlier* sendiri merupakan data yang berada diluar *whiskers* dan biasanya ditampilkan sebagai titik-titik individu pada *boxplot.* Proses *pre-processing* pada data *microarray* bertujuan untuk menghilangkan nilai-nilai non-biologis, sehingga hanya data biologis dan memiliki rata-rata yang seragam saja yang tersisa dalam *boxplot.*

Pada gambar 4 merupakan visualisasi dari nilai intensitas ekspresi gen pasien stroke iskemik saat sebelum dan sesudah *pre-processing*. Dapat dilihat bahwa data intensitas ekspresi gen setelah *pre-processing* lebih komparabel dibanding sebelum *pre-processing*, ditandai dengan variasi *boxplot boxplot* yang rendah (panjang *boxplot* cenderung sama).



Gambar 4. *Boxplot* data sebelum (kiri) dan sesudah (kanan) *pre-processing*

Filtering

Filtering merupakan tahap yang bertujuan untuk menentukan gen-gen yang akan digunakan dalam analisis. Proses ini menggunakan fungsi *nsFilter* (*Non-specified filtering*) yang umum digunakan dalam analisis data *microarray* atau ekspresi gen untuk melakukan pemfilteran pada data. Fungsi *nsFilter* digunakan untuk menghapus fitur (seperti gen atau probe) yang tidak memenuhi kriteria tertentu. Tujuannya adalah untuk membersihkan data dengan menghapus fitur-fitur yang tidak informatif atau yang dapat mengganggu analisis lebih lanjut.

Data hasil *filtering* kemudian akan menjadi *input* untuk *feature selection* menggunakan fungsi *multtest*. *Feature selection* memiliki tujuan untuk mengurangi dimensi dari kumpulan data dengan mengenali *subset* fitur yang paling relevan atau informatif. Oleh karena itu, diharapkan hanya fitur-fitur yang paling signifikan atau penting yang dipertahankan, sedangkan fitur-fitur yang kurang relevan atau redundan dieliminasi. Reduksi dimensi tersebut berpotensi mengurangi kompleksitas data dan meningkatkan kecepatan proses analisis.

Klasifikasi SVM

Setelah tahap *pre-processing* dan *filtering* selesai, diperoleh data yang kemudian akan digunakan sebagai input dalam analisis klasifikasi menggunakan metode SVM. Tahap pertama dalam proses klasifikasi adalah membagi data menjadi data latih dan data uji. Data latih mencakup data dari setiap kelas, yaitu stroke iskemik dan kontrol, yang akan digunakan oleh algoritma SVM untuk mempelajari pola dari masing-masing kelas. Algoritma SVM yang telah dilatih dengan data latih kemudian akan diuji menggunakan data uji untuk mengukur tingkat akurasi dalam memprediksi data baru. Proses ini dikenal sebagai pembelajaran mesin. Data yang digunakan dibagi menjadi 80% untuk data latih dan 20% untuk data uji.

Tabel 2. Pembagian data latih dan data uji

	Data latih	**Data uji**
Rasio	80%	20%
Jumlah	32	8

Pada penilitian ini akan dilakukan percobaan menggunakan kernel-kernel yang terdapat pada algoritma SVM. Kernel-kernel tersebut yaitu *linear, RBF, sigmoid* dan *polynomial*. Selain itu penulis juga melakukan studi simulasi terhadapt nilai *cost optimal* pada masing-masing kernel. Hasil analisis pada tiap kernel dan studi simulasi terhadap nilai *cost optimal* disajikan pada tabel 3.

Tabel 3. Kernel dan *optimal cost*

Kernel	***Optimal Cost***	**Akurasi**
Linear	0,01	73,6%
RBF	1	88,0%
Sigmoid	0,1	0,5%
Polynomial	10	81,8%

Berdasarkan perbandingan nilai akurasi yang diperoleh dari ke empat kernel, diketahui bahwa akurasi terbaik adalah model SVM dengan kernel *RBF* dan *cost* 1. Model SVM dengan kernel *sigmoid* tidak dapat digunakan untuk klasifikasi karena nilai akurasinya yang sangat rendah. Sementara itu, SVM dengan kernel *linear* dan *polynomial* masih tetap dapat digunakan karena nilai akurasinya >70%. Evaluasi dari pembelajaran mesin (*machine learning*) dengan menggunakan kernel RBF pada *cost* 1 dapat dilihat dari hasil uji data latih dan data uji menggunakan *confusion matrix* pada tabel 4 dan 5.

Tabel 4. *Confusion Matrix* pada data latih

Data Aktual	**Data prediksi**		**Total**
	Kontrol	**Stroke Iskemik**	
Kontrol	15	2	17

Stroke iskemik	1	14	15
Total	16	16	32

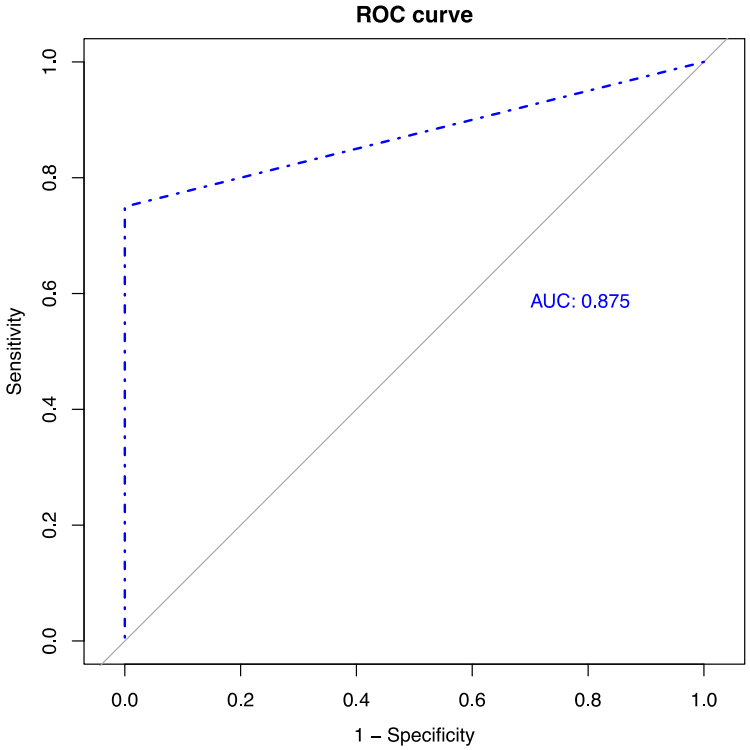
Tabel 5. *Confusion Matrix* pada data uji

Data Aktual	Data prediksi		Total
	Kontrol	Stroke Iskemik	
Kontrol	4	1	5
Stroke iskemik	0	3	3
Total	4	4	8

Beradasrkan tabel 4 dan 5 diketahui, pada data latih terdapat sebanyak 15 sampel data yang diklasifikasikan ke dalam kelas kontrol dan 14 sampel data yang diklasifikasikan kedalam kelas stroke iskemik oleh model, sehingga sisa 3 sampel data diklasifikasikan dengan tidak tepat. Sedangkan pada data uji terdapat 4 sampel data yang diklasifikasikan kedalam kelas kontrol dan 3 sampel data yang diklasifikasikan kedalam kelas stroke iskemik oleh model SVM, sehingga hanya 1 sampel data yang diklasifikasikan dengan tidak tepat. Kemudian dari *confusion matrix* dapat diketahui nilai *accuracy*, *recall*, *precision*, *spesificity* dan *F1-Score* yang disajikan pada tabel 6. Kemudian untuk nilai AUC yang diperoleh yaitu 0,875 dengan grafik ROC secara visual ditampilkan pada gambar 5.

Tabel 6. Metrik evaluasi model SVM pada data latih dan data uji

Metrik	Data prediksi	
	Data latih	Data uji
Accuracy	90,6%	87,5%
Recall	93,7%	100%
Precision	88,2%	80%
Spesificity	87,5%	75%
F1-Score	90,8%	88,8%



Gambar 5. Grafik ROC dari model SVM dengan kernel RBF dan *cost* = 1

Berdasarkan performa dari setiap ukuran metrik yang diperoleh, dapat disimpulkan bahwa model SVM dengan kernel RBF dan *cost* = 1 dapat digunakan untuk memprediksi apakah seorang pasien termasuk kedalam kelas kontrol atau kelas stroke iskemik berdasarkan profil genomic pasien. Selanjutnya, untuk mengetahui gen mana yang berperan sangat penting dalam pengklasifikasian ini, dilakukan ekstrak informasi *variable importance* dari

model yang telah dibuat. Tabel 7 menampilkan urutan *gen importance* disertai dengan variabel lain seperi *id probe*, *gen symbol*, nama gen dan ontologi gen nya.

Tabel 7. Tabel urutan *gene importance*

No	Probe id	Simbol Gen	Nama Gen	Gene Ontology
1	235490_at	TMEM107	transmembrane protein 107	BP, MF, CC
2	218340_s_at	UBA6	ubiquitin like modifier activating enzyme 6	BP, MF, CC
3	209170_s_at	GPM6B	glycoprotein M6B	BP, MF, CC
4	208924_at	RNF11	ring finger protein 11	BP, MF, CC
5	229422_at	NRDC	nardilysin convertase	BP, MF, CC
6	202223_at	STT3A	STT3 oligosaccharyltransferase complex catalytic subunit A	BP, MF, CC
7	205345_at	BARD1	BRCA1 associated RING domain 1	BP, MF, CC
8	208819_at	RAB8A	RAB8A, member RAS oncogene family	BP, MF, CC
9	226273_at	CLCN5	chloride voltage-gated channel 5	BP, MF, CC
10	226733_at	PFKFB2	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2	BP, MF, CC
11	208932_at	PPP4C	protein phosphatase 4 catalytic subunit	BP, MF, CC
12	203861_s_at	ACTN2	actinin alpha 2	BP, MF, CC
13	219861_at	DNAJC17	DnaJ heat shock protein family (Hsp40) member C17	BP, MF, CC
14	223128_at	FOXRED1	FAD dependent oxidoreductase domain containing 1	BP, MF, CC
15	1563872_at	VSTM2B	VSTM2B divergent transcript	NA
16	229365_at	PPP1R3F	protein phosphatase 1 regulatory subunit 3F	BP, MF, CC
17	201747_s_at	SAFB	scaffold attachment factor B	BP, MF, CC
18	210523_at	BMPR1B	bone morphogenetic protein receptor type 1B	BP, MF, CC

5. Kesimpulan

Hasil gambaran dan pengolahan data ekspresi gen PBMCs diketahui pasien penderita stroke iskemik sebanyak 20 pasien control dan 20 pasien lainnya merupakan pasien non control. Dalam proses pengolahan data bioinformatika, terdapat dua tahapan yaitu pre-processing dan filtering. Pre-processing digunakan untuk membersihkan data dari data yang bersifat non biologis, melalui proses background correction, normalization dan sumamrization. Filtering digunakan untuk memilih gen-gen yang signifikan terhadap analisis yang akan dilakukan.

Berdasarkan performa dari setiap ukuran metrik yang diperoleh, dapat disimpulkan bahwa model SVM dengan kernel RBF dan cost = 1 dapat digunakan untuk memprediksi apakah seorang pasien termasuk kedalam kelas kontrol atau kelas stroke iskemik berdasarkan profil genomic pasien. Hasil analisis dan percobaan pada tiap kernel didapatkan nilai akurasi sebesar 88,0% untuk kernel RBF, 73,6% untuk kernel linear, 81,8,5% untuk kernel polynomial dan 0,5% untuk kernel sigmoid. Sehingga berdasarkan penelitian yang dilakukan menggunakan SVM diketahui kernel terbaik untuk klasifikasi pada data ekspresi gen PBMCs stroke iskemik adalah metode SVM dengan kernel RBF dan nilai optimal cost yang digunakan yaitu 1.

6. Daftar Pustaka

[1] S. C. Smeltzer and B. G. Bare, *Brunner and Suddarth's Textbook of Medical-Surgical Nursing*, Philadelphia: Lippincott-Raven Publisher, 1996.

[2] K. and R. D. Saraswati, "Transisi Epidemiologi Stroke sebagai Penyebab Kematian pada Semua Kelompok Usia di Indonesia," in *Prosiding Seminar Nasional Riset Kedokteran (SENSORIK)*, Jakarta, 2021.

[3] Y. Haiga, I. P. P. Salman and S. Wahyuni, "Perbedaan Diagnosis Stroke Iskemik dan Stroke Hemoragik dengan Hasil Transcranial Doppler di RSUP Dr. M. Djamil Padang," *Scientific Journal*, vol. 1, no. 5, 2022.

[4] S. M. Thampi, Bioinformatics, Kerala: LBS College of Engineering, 2017.

[5] H. S. and T. S. Famuji, "Proses Implementasi Bioinformatika Pada Digitalisasi Data Genetika Manusia," *Jurnal SIMETRIS*, vol. 14, no. 1, 2023.

[6] A. A. Parikesit, "Kontribusi Aplikasi Medis dari Ilmu Bioinformatika Berdasarkan Perkembangan Pembelajaran Mesin (Machine Learning) Terbaru," *Cermin Dunia Kedokteran*, vol. 45, no. 9, 2018.

[7] L. Zimmermann, "Completly Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core," *Journal of Molecular Biology*, vol. 430, no. 15, 2018.

[8] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan and X. Gao, "Deep Learning in Bioinformatics: Introduction, Application, and Prespective in the Big Data Era," *Methods*, vol. 166, no. 10, 2019.

[9] A. M. Puspita, D. E. Ratnawati and A. W. Widodo, "Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 2, 2018.

[10] S. H. Wibowo and R. Toyib, "Support Vector Machine Method for Recognizing Patterns in Signatures," *Jurnal Media Infotama*, vol. 18, no. 2, 2022.

[11] P. Fremmuzar and A. Baita, "Uji Kernel SVM dalam Analisis Sentimen Terhadap Layanan Telkomsel di Media Sosial Twitter," *Komputika: Jurnal Sistem Komputer*, vol. 12, no. 2, 2023.

[12] P. N. Andono and E. H. Rachmawanto, "Evaluasi Ekstraksi Fitur GLCM dan LBP Menggunakan MultikernelSVM untuk Klasifikasi Batik," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, p. 4, 2020.

[13] R. Permatasari and A. Wibowo, "Implementation of support Vector Machine - Recursive Feature Elimination for MicroRNA Selection in Breast Cancer Classification," *Jurnal EECCiS*, vol. 14, no. 1, 2020.

[14] L. Syafa'ah, Z. Zulfatman, I. Pakaya and M. Lestandy, "Comparison of Machine Learning Classification Methods in Hepatitis C Virus," *Jurnal Online Informatika (JOIN)*, vol. 6, no. 1, 2021.

[15] D. V. Carreras, J. Alcaraz and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Computers & Operations Research*, vol. 152, 2023.

[16] R. Chairunisa, A. and W. Astuti, "Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray," *Jurnal RESTI*, vol. 4, no. 5, 202.