Studi QSAR pada Prediksi Inhibitor DPP-IV sebagai Agen Anti Diabetes Menggunakan Metode Algoritma Genetika-*Support Vector Machine*

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi S1 Informatika

Fakultas Informatika

Universitas Telkom

1301164049

Attariq Muhammad Kasfilla

Program Studi Sarjana Teknik Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2021

LEMBAR PENGESAHAN

Studi QSAR pada Prediksi Inhibitor DPP-IV sebagai Agen Anti Diabetes Menggunakan Metode Algoritma Genetika-*Support Vector Machine*

QSAR Study On Predicting DPP-IV Inhibitors as Anti-Diabetic Agent using Genetic Algortihm-*Support Vector Machine*

NIM :1301164049

Attariq Muhammad Kasfilla

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana S1 Informatika

Fakultas Informatika

Universitas Telkom

Bandung, 11/Febuari/2021

Menyetujui

Pembimbing I,	Pembimbing II,
Isman Kurniawan, S.Pd., M.Si., M.Sc., Ph.D.	Nurul Ikhsan, S.Si., M.Si., M.Sc., Ph.D.
NIP : 1587006	NIP : 14870044

Ketua Program Studi
Sarjana S1 Informatika,

Niken Dwi Wahyu Cahyani, ST., M.Kom., Ph.D
NIP: 00750052

LEMBAR PERNYATAAN

Dengan ini saya, Attariq Muhammad Kasfilla, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul Studi QSAR pada Prediksi Inhibitor DPP-IV sebagai Agen Anti Diabetes Menggunakan Metode Algoritma Genetika-*Support Vector Machine* beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 11/Febuari/2021
Yang Menyatakan

Attariq Muhammad Kasfilla

Studi QSAR pada Prediksi Inhibitor DPP-IV sebagai Agen Anti Diabetes Menggunakan Metode Algoritma Genetika-*Support Vector Machine*

Attariq Muhammad Kasfilla[1], Isman Kurniawan [2], Nurul Ikhsan[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung,
[1]attrqmksflla@students.telkomuniversity.ac.id, [2]ismankrn@telkomuniversity.ac.id,,
[3]ikhsan@telkomuniversity.ac.id

Abstrak

Diabetes melitus merupakan penyakit degeneratif utama di abad ke-21 yang menyebabkan hampir 95% orang dewasa didiagnosis penyakit diabetes tipe II. Salah satu enzim yang bertanggung jawab terhadap diabetes tipe II adalah Dipeptidyl peptidase-IV (DPP IV). Akibat terbatasnya jumlah inhibitor untuk diabetes tipe II, ada kebutuhan mendesak untuk mengembangkan tambahan inhibitor DPP IV yang baru. Seiringnya perkembangan teknologi, banyak penelitian yang terlibat dalam penemuan dan optimalisasi inhibitor DPP IV baru sebagai pengobatan diabetes tipe II dengan menggunakan metode *quantitative structure-activity relationship* (QSAR). Banyak penelitian telah menemukan atau mencoba mencari salah satu obat penyakit diabetes melitus dengan menggunakan metode ini. Pada penelitian ini, penulis bertujuan membuat model untuk memprediksi aktivitas anti diabetes tipe II. Tahapan penelitian ini dibagi menjadi 2, yaitu seleksi fitur dan membangun model prediksi. Tahap seleksi fitur dilakukan dengan menggunakan metode algoritma genetika (AG), dan tahap model prediksi menggunakan metode *support vector machine* (SVM). Berdasarkan hasil yang didapatkan, penulis menemukan model prediksi terbaik pada kernel RBF dibandingkan dengan kernel lainnya. Hal ini ditunjukan dengan akurasi yang diperoleh sebesar 0.9869, *precision* 0.9745, *recall* 1.0, dan *F1-Score* 0.9871.

Kata kunci : Diabetes melitus, dipeptidyl peptidase-IV, QSAR, algoritma genetika, *Support Vector Machine*

Abstract

Diabetes mellitus is a major degenerative disease in the 21st century which causes nearly 95% of adults to be diagnosed with type II diabetes. One of the enzymes responsible for type II diabetes is Dipeptidyl peptidase-IV (DPP IV). Due to the limited number of inhibitors for type II diabetes, there is an urgent need to develop additional new DPP IV inhibitors. Along with technological developments, a lot of research is involved in the discovery and optimization of new DPP IV inhibitors as a treatment for type II diabetes using the quantitative structure-activity relationship (QSAR) method. Many studies have found or tried to find a cure for diabetes mellitus using this method. In this study, the authors aimed to create a model to predict anti-diabetes type II activity. The stages of this research are divided into 2, namely feature selection and building a prediction model. The feature selection stage is carried out using the genetic algorithm (GA) method, and the prediction model stage uses the support vector machine (SVM) method. Based on the results obtained, the authors found the best prediction model in the RBF kernel compared to other kernels. This is shown by the accuracy obtained by 0.9869, precision 0.9745, recall 1.0, and F1-Score 0.9871.

Keywords: Diabetes melitus, dipeptidyl peptidase-IV, QSAR, Genetic Algorithm, Support Vector Machine

A. Pendahuluan

Diabetes melitus adalah penyakit yang diakibatkan kadar glukosa (gula) darah lebih tinggi dari normal[1], [2]. Penyakit ini merupakan salah satu penyakit degeneratif utama di abad ke-21[3]. Ada tiga jenis utama diabetes melitus, yaitu tipe I, tipe II, dan diabetes gestasional[3]. Obesitas adalah faktor utama penyebab diabetes[1]. Adapun faktor lain penyebab diabetes seperti resistensi insulin, produksi insulin yang tidak mencukupi, hiperglikemia kronis, peningkatan produksi glukosa hepatik, atau intoleransi glukosa[3]. Hampir 95% diabetes yang didiagnosis pada orang dewasa adalah diabetes tipe II[1]. Tanpa pengobatan, diabetes akan "melemahkan" fungsi tubuh secara bertahap yang dapat mengakibatkan amputasi anggota tubuh, kebutaan, penyakit hati berlemak, penyakit ginjal serta kematian dini[1].

Salah satu enzim yang bertanggung jawab terhadap diabetes tipe II adalah Dipeptidyl peptidase-IV (DPP IV) [2], [4]. Perhatian intensif telah diberikan kepada DPP-IV sebagai target penting untuk pengobatan diabetes tipe II[3]. Ada bukti yang menunjukkan bahwa penghambatan DPP-IV sebagai strategi untuk meningkatkan status glikemik lebih efektif pada diabetes tipe II[2], [5]. Oleh karena itu, para ilmuwan bidang kimia medis mengembangkan senyawa untuk menghambat DPP-IV, sehingga mengarah pada

peningkatan sekresi insulin, dan dengan demikian memperbaiki hiperglikemia pada diabetes tipe II[3]. Banyak kelompok penelitian, yang terlibat dalam penemuan dan optimalisasi inhibitor DPP IV baru sebagai pengobatan diabetes tipe II[6].

Akibat terbatasnya jumlah inhibitor untuk diabetes tipe II yang sukses di pasaran, ada kebutuhan mendesak untuk mengembangkan tambahan inhibitor DPP-IV yang efektif berguna secara terapi[6]. Pendekatan desain obat yang rasional telah menjadi komponen dari setiap paradigma penemuan obat[7]. Metode QSAR dapat menjadi alternatif dalam proses desain obat maupun mempelajari hubungan antara struktur molekul dengan aktivitas biologisnya. Terdapat beberapa kelebihan dari metode QSAR, salah satunya untuk mengurangi biaya dan risiko dalam industri obat-obatan[4]. Studi pemodelan QSAR telah dilakukan dan diusulkan pada beberapa data set inhibitor DPP-IV. Penggunaan metode QSAR dipilih karena tingkat ketelitiannya cukup tinggi sedangkan waktu perhitungannya rendah dalam pengembangan obat[3], [4]. Oleh karena itu, dalam beberapa tahun terakhir, inhibitor yang didasarkan pada berbagai perancah telah dijelaskan dalam literatur dan beberapa bahkan telah menjalani berbagai tahap uji klinis seperti vildagliptin sitagliptin, saxagliptin sebagai obat anti diabetes yang ampuh[2].

Penelitian ini bertujuan untuk memprediksi aktivitas inhibitor DPP-IV dan mengetahui performansi kombinasi dari dua metode algoritma genetika dengan *support vector machine* (SVM). Tahapan penelitian dibagi menjadi 2, yaitu tahap seleksi fitur dan membangun model prediksi. Metode yang digunakan dalam tahap seleksi fitur adalah algoritma genetika (AG). AG dipilih karena metode tersebut jumlah parameternya sangat besar, dan dapat mencari paralel dari populasi titik. Sehingga, AG memiliki kemampuan untuk menghindari dari terjebaknya dalam solusi optimal lokal seperti metode tradisional, yang mencari dari satu titik. Adapun tahap model prediksi dibangun menggunakan metode SVM. SVM dipilih karena metode tersebut efektif apabila digunakan dalam kasus dengan jumlah dimensi lebih besar dari jumlah sampel dan dalam ruang dimensi yang tinggi[8].

B. Material dan Metode

1. Studi Literatur

Desain obat sangat terkait dengan pemodelan struktur-aktivitas hubungan kuantitatif (QSAR). Beberapa prosedur pemodelan seperti regresi dan klasifikasi dapat digunakan dalam membangun model QSAR yang efisien untuk desain obat [3]. Metode QSAR sudah banyak digunakan untuk memprediksi aktivitas agen anti-diabetes. Pada tahun 2009 U. Saqib dan M.I. Siddiqi melakukan studi analisis 3D CoMFA dan CoMSIA QSAR untuk memprediksi aktivitas antidiabetik pada 45 turunan triazolopiperazine amida sebagai inhibitor DPP-IV. Informasi yang diperoleh dari kedua model, memberikan korelasi yang cukup signifikan dari berbagai deskriptor, sehingga inhibitor yang lebih kuat dapat dirancang dan disintesis untuk analisis aktivitas anti-diabetes sebagai obat anti-diabetes dari analog berbasis triazolopiperazine amida [2].

Pada tahun 2013 Anjar Purba Asmara dan rekannya menggunakan metode semiempirik AM 1 (Austin Model 1) untuk memprediksi senyawa turunan triazolopiperazin amida, sebanyak 22 senyawa digunakan sebagai inhibitor enzim DPP IV. Mereka menyimpulkan metode semiempirik AM1 menghasilkan model senyawa turunan triazolopiperazin amida yang baik [4][4]. Selanjutnya pada tahun 2018 Marcos Lorca dengan rekan-rekannya, melakukan penelitian QSAR dengan menggunakan metode CoMFA and CoMSIA dimana reseptor β3 adrenergik sebagai target potensial untuk pengobatan beberapa patologi seperti diabetes, obesitas, dll. Dalam 10 tahun terakhir, hanya ada dua laporan penelitian hubungan struktur-aktivitas kuantitatif (QSAR) tentang senyawa selektif untuk β3-AR. Mereka menyimpulkan model CoMFA dan CoMSIA masing-masing menyajikan validasi internal yang baik [9].

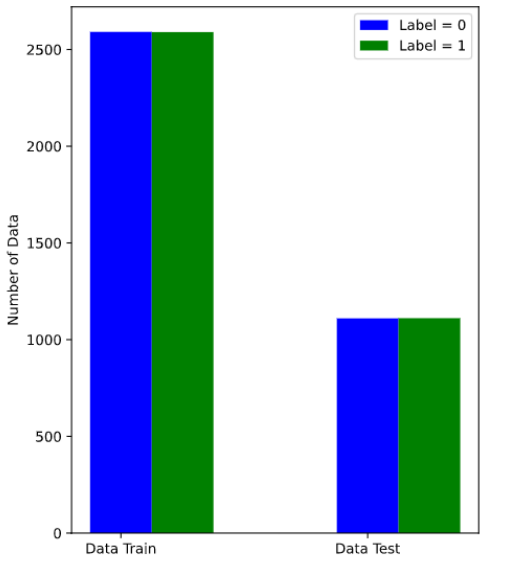
Pada tahun 2019 A.M. Al-Fakih dengan rekannya melakukan penelitian QSAR dengan target DPP-IV. Mereka mengusulkan sebanyak 134 inhibitor diteliti dengan metode algoritma pencarian gravitasi biner yang bervariasi waktu (TVBGSA). Mereka mendapatkan kesimpulan kekuatan prediksi yang lebih baik dari model TVBGSA dibandingkan dengan BGSA. Penelitian mereka mengusulkan TVBGSA sebagai pendekatan pemodelan untuk membangun QSAR yang andal dan kuat untuk memprediksi aktivitas antidiabetes dari inhibitor DPP-IV sebelum merancang dan melakukan sintesis sintesis inhibitor DPP-IV baru. Meskipun hasil TVBGSA menghasilkan kinerja yang jauh lebih baik, TVBGSA memiliki keterbatasan. Kinerjanya sepenuhnya tergantung pada nilai maksimum dari parameter kontrol [3].

2. Dataset

Pada penelitian ini, data yang digunakan merupakan data senyawa penghambat DPP-IV. Data diambil dari ChemBL[10] yang berisi 4.757 senyawa. Data dibagi ke dalam dua sub-data menggunakan IC50 sebagai kriterianya. Nilai IC50 digunakan sebagai parameter, untuk menunjukan kekuatan zat penghambat DPP-IV. Sub-data pertama merupakan data yang dipilih dengan nilai IC50 <= 10µM dan sub-data kedua merupakan data yang nilai IC50 => 100µM. Untuk data yang tidak berada di dua kondisi tersebut, data akan dihilangkan. Sub-data pertama diberi label 1 yang merupakan senyawa aktif dengan jumlah data sebanyak 3701, sedangkan untuk sub-data kedua diberi label 0 yang merupakan senyawa putative dengan jumlah data sebanyak 113. Karena jumlah sub-data yang tidak seimbang, dilakukan perhitungan untuk menyeimbangkan jumlah data. Proses ini dilakukan dengan membuat beberapa pengelompokkan pada kumpulan senyawa dalam jumlah besar dan kelompok tersebut diseleksi. Kelompok yang digunakan merupakan kelompok yang hanya memiliki senyawa putative saja. Total data menjadi 7.402 data dengan masing-masing jumlah sub-data sebanyak 3701.

Selanjutnya, untuk mendapatkan nilai dari tiap deskriptor molekuler dibutuhkan struktur masing-masing senyawa. Deksriptor molekuler diperoleh dengan notasi *simplified molecular input line-entry system* (SMILES) yang akan dihitung menggunakan PaDEL-Descriptor[11]–[13]. Hasil yang didapatkan dari PaDEL-Descriptor merupakan kumpulan fitur deksriptor molekuler sebanyak 627 fitur. Data yang telah dihitung, selanjutnya data secara acak dipilih sehingga terbagi menjadi dua set, 70% sebagai data set latih dan 30% sebagai data set uji.

Jumlah masing-masing data yaitu 5.181 data latih dan 2.221 data uji. Adapun jumlah data dengan senyawa aktif dan tidak aktif untuk data latih dan data uji dapat dilihat pada Gambar 1. Jumlah masing-masing inhibitor pada data latih yaitu sebanyak 2.591 senyawa aktif dan 2.590 senyawa tidak aktif. Sedangkan pada data uji jumlah senyawa aktif yaitu 1.111 dan 1.110 untuk senyawa tidak aktif.



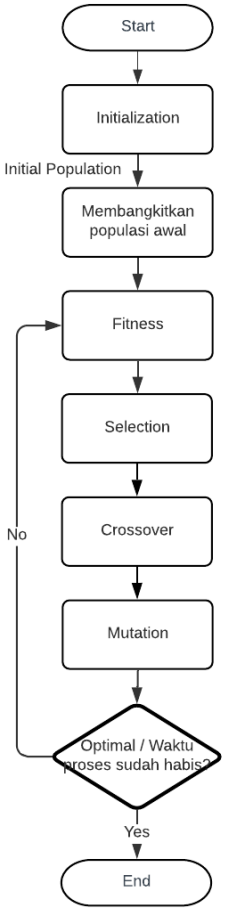
Gambar 1. Data Managing

3. Seleksi Fitur

Di dalam seleksi fitur, penulis menggunakan *Pearson coefficient correlation* (PCC) dan Algoritma Genetika (AG). PCC merupakan metode yang berfungsi untuk menghitung korelasi antar semua fitur secara analisis statistik dan arah hubungan linier antara dua variabel acak[14]. Setelah data dihitung, kemudian data diurutkan dan diambil 100 fitur dengan korelasi terbesar terhadap target. Lalu selanjutnya data yang telah diurutkan akan dimasukan kedalam metode AG.

AG merupakan teknik pencarian dan optimasi acak yang dipandu oleh prinsip-prinsip meniru evolusi biologis dalam sistem genetik alami yang dapat digunakan untuk menyelesaikan masalah dan memodelkan sistem evolusi[15]. AG sering disebut metode komputasi evolusioner yang digunakan

dalam aspek adaptif komputasi-pencarian, optimasi, pembelajaran mesin, penyesuaian parameter, dll[13], [15]. Algoritma ini mempertahankan dan memanipulasi suatu populasi solusi dan mengimplementasikan pencarian mereka untuk solusi yang lebih baik agar digunakan, seperti halnya algoritma pencarian lainnya mencari solusi potensial agar memecahkan masalah[11].



Gambar 2. Flow Chart Algoritma Genetika

Menurut [11], [15], terdapat 7 langkah dasar di dalam algoritma genetika, yaitu Representation, Population, Fitness, Selection, Crossover, Mutation, dan Termination. *Representation* dalam AG digunakan untuk mencari solusi masalah. Solusi potensial untuk masalah harus dikodekan sebagai rangkaian karakter. Setiap string adalah urutan bilangan real yang mewakili pusat cluster K. Untuk ruang dimensi-N, panjang kromosom adalah N * K, di mana "posisi N pertama (atau, gen) mewakili dimensi N dari" pusat klaster pertama, posisi N berikutnya mewakili kluster kedua pusat, dan seterusnya [11]. Setelah menggambarkan solusi masalah, selanjutnya *population* yang merupakan Pusat-pusat klaster K yang disandikan di setiap kromosom diinisialisasi ke K yang dipilih secara acak dari set data. Proses ini diulangi untuk setiap kromosom P dalam populasi, di mana P adalah ukuran populasi[15].

Untuk mendapatkan nilai dari setiap kromosom pada populasi, *fitness* bertugas untuk menemukan string yang sangat baik. Kebaikan suatu string direpresentasikan dalam AG oleh beberapa fungsi string, yang penulis sebut fungsi objektif, kuantitas yang akan dioptimalkan. Fungsi lain yang dibutuhkan disebut fungsi kebugaran. Ini adalah fungsi monoton fungsi obyektif, dan inilah yang menentukan bagaimana string akan dikalikan atau dibuang selama algoritma[11], [15]. Data-data yang telah dihitung nilainya, kemudian diseleksi atau di dalam AG adalah *selection*, yaitu memilih kromosom dari populasi sehingga kromosom yang lebih baik lebih mungkin untuk dipilih. Seleksi diterapkan berulang-ulang, ia hanya akan membuat proporsi kromosom meningkat apabila kromosom

memang lebih baik di dalam populasi, dan mengurangi jumlah kromosom yang kurang baik[11], [13], [15].

Selanjutnya *crossover* yaitu operasi (proses) probabilistik penggabungan informasi dari sepasang orang tua (induk) menghasilkan dua (atau beberapa) keturunan. Setiap keturunan mengandung informasi dari string masing-masing 'induk' yang diharapkan menghasilkan solusi 'keturunan' yang lebih baik[13], [15]. Kemudian, *mutation* adalah proses yang mengubah secara acak nilai setiap elemen kromosom sesuai dengan probabilitas mutasi karakter. Untuk AG yang berkode nyata, operator mutasi meliputi mutasi seragam, mutasi non-seragam, mutasi multi-non-seragam, mutasi batas, dll[11]. Terakhir *termination* : String terbaik yang terlihat hingga generasi terakhir memberikan solusi untuk masalah pengelompokan. Penulis telah menerapkan elitisme pada setiap generasi dengan mempertahankan string terbaik yang terlihat hingga generasi itu di lokasi di luar populasi. Dengan demikian pada penghentian, lokasi ini berisi pusat - pusat dari "cluster akhir. Bagian berikutnya memberikan hasil implementasi algoritma AG-clustering, bersama dengan perbandingannya dengan kinerja algoritma K-means untuk beberapa arti "set data kehidupan nyata dan nyata[16], [17].

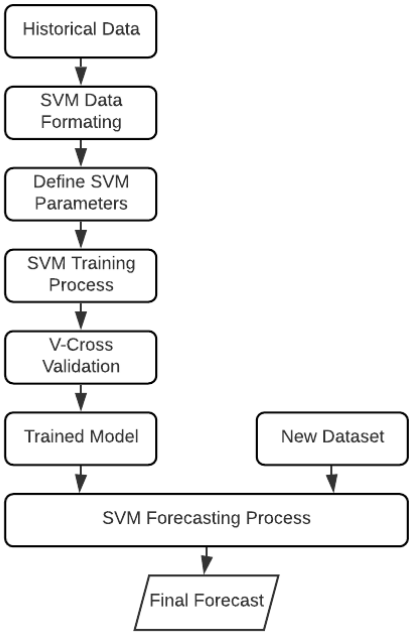
Untuk menangani faktor acak AG, penulis menjalankan 20 penghitungan di sistem dan membandingkan setiap nilai MSE yang dihasilkan. Nilai MSE terendah dipilih untuk dijadikan sebagai acuan dalam membentuk kombinasi terbaik. Di dalam seleksi fitur terdapat beberapa parameter seperti Populasi, *Fitness*, *Crossover*, *Mutation*, Iterasi, *Parent Selection* dan *Survivor Selection* dengan nilai seperti yang terlihat pada Tabel 1 sebagai berikut:

Tabel 1. Parameter AG

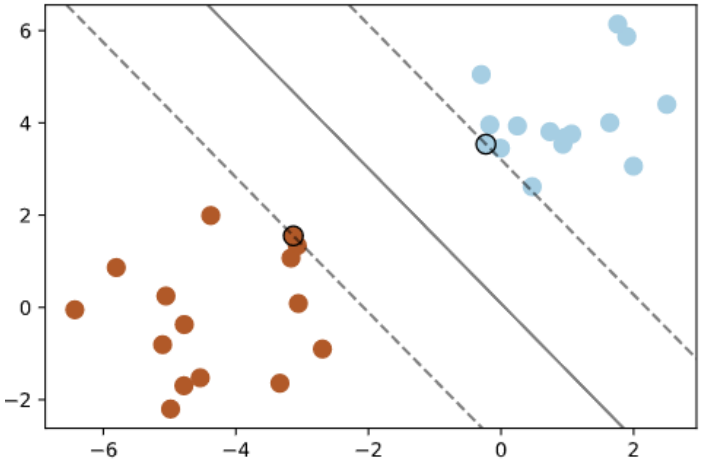
Parameter	**Nilai**
Populasi	10
Fitness	MSE $1/(Logloss + 1e - 10)$
Crossover	1 Titik
Mutation	0,01%
Iterasi	100x
Parent Selection	*Roulette Wheel Selection*
Survivor Selection	*Steady-State (Fitness-based selection)*

4. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah alat atau metode yang digunakan untuk memecahkan masalah pengenalan klasifikasi dan regresi[16]. Selama beberapa tahun terakhir, SVM telah menarik banyak peneliti dari jaringan saraf dan komunitas pemrograman matematika. Alasan utama menggunakan metode ini adalah kemampuannya untuk memberikan kinerja generalisasi yang sangat baik. SVM juga telah terbukti bermanfaat untuk beberapa aplikasi di dunia nyata[18]–[20]. Algoritma ini telah disampaikan pada tahun 1995 oleh Vladimir Vapnik dan rekannya, meskipun dasar untuk SVM telah ada pada tahun enam puluhan di Rusia [19]. Tujuan dari SVM untuk menemukan fungsi f(x) yang memiliki paling banyak penyimpangan ε dari target yi yang sebenarnya diperoleh untuk semua data pelatihan, dan pada saat yang sama serata mungkin[19]. Gambar 3 merupakan *flowchart* dari SVM.



Gambar 3 *Flowchart* SVM



Gambar 4. Hyperplane

$$W . X + b = 0 \qquad (1)$$

Tujuan sederhana dari SVM adalah mencari fungsi pemisah (*hyperplane*) seperti pada Gambar 4, di mana *hyperplane* tersebut berfungsi untuk memisahkan dua set data dari dua kelas yang berbeda, *hyperplane* dapat dipasang dengan persamaan (1). Dengan W adalah sebagai bobot, X merupakan koordinat dari titik data tersebut, dan b merupakan bias. *Hyperplane* terbaik adalah *hyperplane* yang terletak di tengah-tengah antara dua set obyek dari dua kelas atau ekuivalen dengan memaksimalkan margin atau jarak antara dua set obyek dari kelas yang berbeda. Di dalam SVM juga terdapat beberapa kernel, seperti Linear, Polynomial, dan RBF yang disajikan seperti pada Tabel 2.

Tabel 2. Persamaan Kernel

Kernel	Persamaan
Linear	$K(x_i, x_j) = (x_i, x_j) + c$

Polynomial	$\mathrm{K}(x_i, x_j) = (x_i, x_j + c)^d$
Gaussian Radial Basic Function	$\mathrm{K}(x_i, x_j) = \exp(-\frac{\lvert x_i, x_j \rvert^2}{2\sigma^2})$

5. Validasi Model

Model divalidasi menggunakan beberapa metrik yang berasal dari *Confusion Matrix*[21], [22]. Pada penelitian ini, *True Positive* (TP) merupakan jumlah data positif yang terklasifikasi dengan benar sedangkan *True Negative* (TN) adalah jumlah data negatif yang terklasifikasi dengan benar oleh system. Adapun *False Negative* (FN), yaitu jumlah data negatif namun terklasifikasi salah dan *False Positive* (FP), yaitu jumlah data positif namun terklasifikasi salah oleh sistem. *Confusion matrix* dapat dilihat pada Tabel 3 sebagai berikut.

Tabel 3. Confusion Matrix

Actual Condition	Predicted Negative	Predicted Positive
Negative	TN	FP
Positive	FN	TP

Dari *confusion matrix* tersebut, dapat dihitung beberapa parameter validasi seperti akurasi (Q), *precision* (PR), *recall* (RE), *F1-score* dan *Area Under the Curve* (AUC). Validasi model dilakukan dengan menghitung parameter statistik terhadap data latih dan data uji sebagaimana yang ditunjukkan pada persamaan berikut:

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \quad (1)$$

$$PR = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

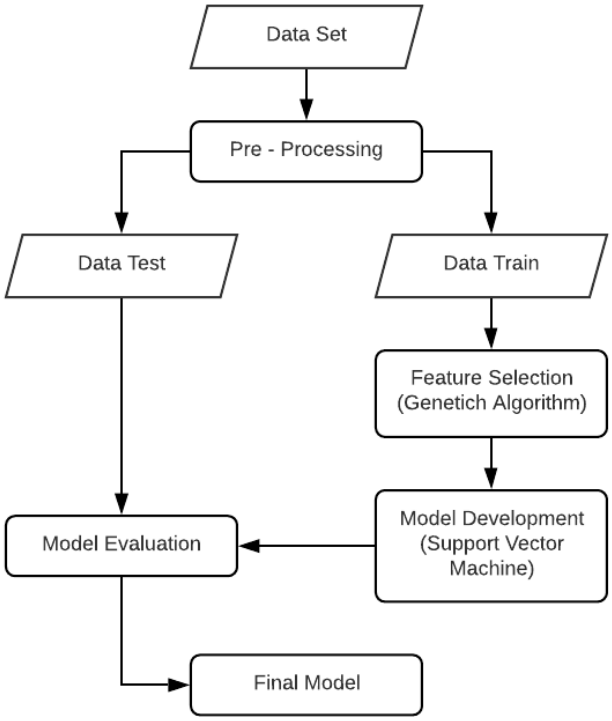
$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

$$F1 = 2 \times \frac{\mathrm{PR} \times \mathrm{RE}}{\mathrm{PR} + \mathrm{SE}} \quad (4)$$

Akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar. *Precision* menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar. *Recall* menunjukkan berapa persen data kategori positif yang terklasifikasikan dengan benar. *F1-score* dapat diartikan sebagai rata-rata tertimbang dari *precision* dan *recall*. AUC untuk mengevaluasi kualitas kinerja pengklasifikasi

6. Diagram Alir Sistem

Diagram alir menunjukkan proses yang dilakukan pada penelitian ini pada gambar 5:



Gambar 5 Flowchart Sistem

Penelitian ini dimulai dari mengumpulkan *dataset* yang diperlukan untuk kegiatan penelitian, pada tahap *preprocessing* data tersebut terbagi menjadi 2 bagian data, sebagian menjadi *train data* dan sebagian lainnya menjadi *test data*. Pada data *train*, data dilakukan *feature selection* dan selanjutnya dilakukan *model development* menggunakan metode *Support Vector Machine*. Pada tahap ketiga yaitu *model development* menggunakan metode *Support Vector Machine* kemudian *data test* akan di model evaluasi yang akan menghasilkan *final model* untuk digunakan pada studi QSAR

C. Hasil dan Pembahasan

1. Seleksi Fitur

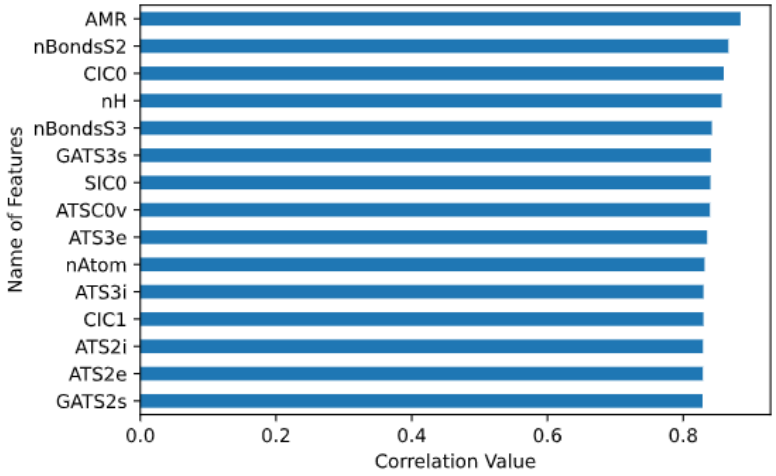
Setelah data diproses menggunakan *pandas correlation pearson*, didapat 100 fitur dengan korelasi terbesar terhadap target untuk digunakan dalam membangun model prediksi. Selanjutnya, data yang telah diurutkan kemudian diolah menggunakan metode AG dan menghasilkan 6 jenis model yang dapat dilihat pada Tabel 4 berikut.

Tabel 4. Hasil Seleksi Fitur

Jumlah Deskriptor	Nama Deskriptor	Nilai MSE Terkecil
5	'nBondsS2', 'ATS4i', 'GATS5s', 'TIC4', 'AATS1i'	0.482954
6	'AMR', 'ATS6i', 'TIC5', 'GATS6s', 'GATS4s', 'ATS7v'	0.476292
7	'Sse', 'ATS2v', 'ATS5p', 'GATS7s', 'McGowan_Volume', 'ZMIC1', 'WTPT-1'	0.469631
8	'GATS3s', 'ATS2v', 'ATS0v', 'GATS8s', 'SpDiam_Dt', 'SIC2', 'Kier1', 'WTPT-1'	0.466299
9	'CIC0', 'nBondsS3', 'AMR', 'IC0', 'GATS5s', 'SIC2', 'ATS2i', 'AMW', 'MW'	0.436324

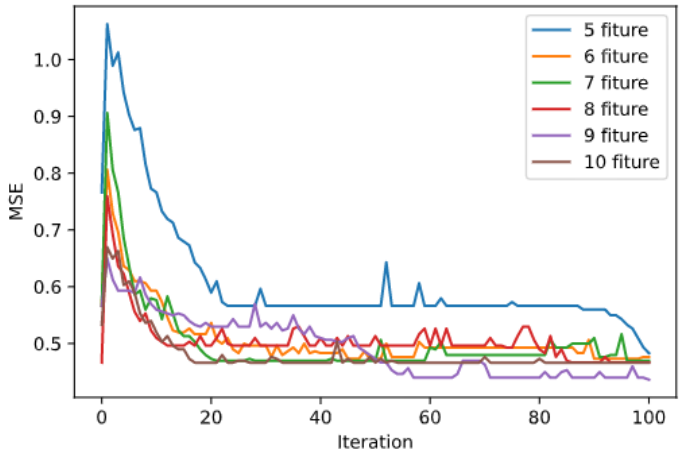
10	'AMR', 'SIC0', 'nC', 'Sv', 'BIC1', 'GATS4s', 'ATS0v', 'GATS8s', 'ATS8p', 'CrippenLogP'	0.466298

Gambar 6 adalah 15 fitur dengan korelasi terbesar terhadap target dari 100 data yang telah dihitung menggunakan pearson correlation coefficient.



Gambar 6. Feature Importance

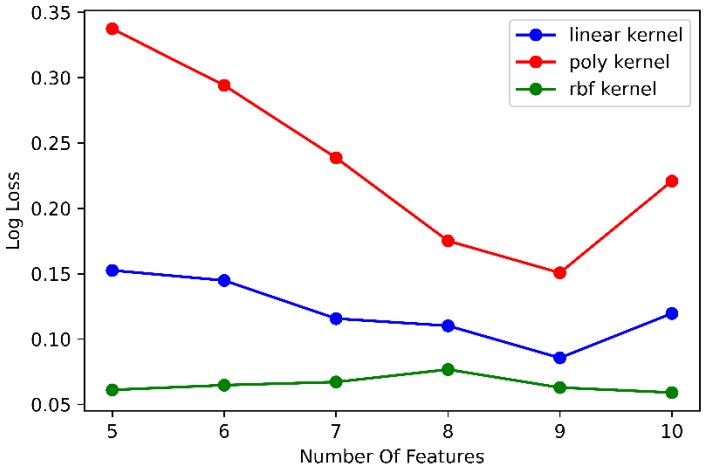
Pada Gambar 7 menunjukan performa seleksi fitur, setiap iterasi yang dilakukan menghasikan nilai MSE untuk 6 jenis model



Gambar 7. MSE vs Iteration

2. Validasi Model

Pembangunan model dilakukan dengan cara melatih setiap data latih berdasrkan 6 jenis kombinasi deskriptor yang telah didapatkan dari seleksi fitur (AG). Deskriptor-deskriptor tersebut selanjutnya diprediksi menggunakan 3 jenis kernel (linear, poly, RBF) untuk mendapatkan nilai *log loss* (besarnya eror) dan menentukan data terbaik untuk tiap jenis kernel berdasarkan nilai *log loss* terendah. Pada Gambar 8, dapat dilihat bahwa setiap jenis kernel memiliki datanya masing-masing. Kernel linear dan poly dengan 9 jenis kombinasi deskriptor serta RBF dengan 10 jenis deskriptor memiliki nilai *log loss* terbaik sebesar 0.085, 0.150, 0.058 secara berturut-turut, seperti yang terlihat pada Tabel 5.



Gambar 8. Log Loss

Tabel 5. Log Loss Value

Jumlah Deskriptor	Nilai Log Loss		
	Linear	**Poly**	**RBF**
5	0.153	0.337	0.061
6	0.145	0.294	0.064
7	0.115	0.238	0.067
8	0.112	0.176	0.076
9	**0.085**	**0.150**	0.062
10	0.119	0.219	**0.058**

Setelah mendapatkan hasil untuk setiap jenis kernel, selanjutnya model digunakan pada data latih dan data uji untuk mendapatkan hasil validasi model. Hasil dari validasi model disajikan seperti pada Tabel 6 sebagai berikut.

Tabel 6 Ringkasan Hasil Validasi

Kernel / Data	TP	FP	TN	FN	Q	PR	RE	F1	AUC
Train									
Linear	2575	84	2507	15	0.9808	0.9684	0.9942	0.9811	0.9808
Poly	2584	78	2513	6	**0.9837**	**0.9706**	0.9976	**0.9840**	**0.9837**
RBF	2588	84	2507	2	0.9834	0.9685	**0.9992**	0.9836	0.9834
Test									
Linear	1107	29	1081	4	0.9851	0.9744	0.9963	0.9853	0.9851
Poly	1109	27	1083	2	**0.9869**	**0.9762**	0.9981	0.9870	**0.9869**
RBF	1111	29	1081	0	**0.9869**	0.9745	**1.0**	**0.9871**	**0.9869**

Dari table 6 kita dapat melihat rata-rata akurasi masing-masing data set sebesar 0.9826 data latih sedangkan untuk data uji sebesar 0.9863. Kernel terbaik untuk memprediksi turunan senyawa DPP-IV adalah kernel RBF dengan akurasi sebesar 0.9869, *precision* 0.9745, *recall* 1.0, dan *F1-Score* 0.9871 hasil dari data uji. Di urutan selanjutnya adalah kernel poly lalu di ikuti oleh kernel linear dengan akurasi sebesar 0.9851, *precision* 0.9744, *recall* 0.9963, dan *F1-Score* 0.9853. Berdasarkan hasil dari data latih dan data uji, setiap kernel menunjukan konsistensi dalam memprediksi senyawa turunan DPP-IV

D. Kesimpulan

Senyawa inhibitor DPP-IV diketahui membawa kandidat obat anti diabetes baru menurut beberapa studi QSAR. Pada penelitian ini tahap seleksi fitur dilakukan dengan menghitung 100 korelasi tertinggi masing-masing deskriptor terhadap pIC50 menggunakan PCC dan menggunakan metode Algortima Genetika (AG) untuk mendapatkan kombinasi deskriptor terbaik dari dataset tersebut. Model prediksi berhasil dibangun menggunakan metode AG dan SVM. AG berhasil mendapatkan kombinasi fitur yang baik dan SVM dapat digunakan untuk memprediksi aktifitas inhibitor DPP-IV dengan akurasi rata-rata sebesar 0.983 untuk data latih dan 0.986 untuk data uji. Hasil juga menunjukan bahwa kernel RBF merupakan kernel terbaik dalam memprediksi senyawa DPP-IV dengan akurasi sebesar 0.9869, *precision* 0.9745, *recall* 1.0, dan *F1-Score* 0.9871.

Daftar Pustaka

[1] UCLA Center for Health Policy Research, “Health Impact of Diabetes,” *Public Health Advocacy*, no. May, 2014.

[2] U. Saqib and M. I. Siddiqi, “3D-QSAR studies on triazolopiperazine amide inhibitors of dipeptidyl peptidase-IV as anti-diabetic agents,” *SAR and QSAR in Environmental Research*, vol. 20, no. 5–6, pp. 519–535, 2009, doi: 10.1080/10629360903278677.

[3] A. M. Al-Fakih, Z. Y. Algamal, M. H. Lee, M. Aziz, and H. T. M. Ali, “A QSAR model for predicting antidiabetic activity of dipeptidyl peptidase-IV inhibitors by enhanced binary gravitational search algorithm,” *SAR and QSAR in Environmental Research*, vol. 30, no. 6, pp. 403–416, 2019, doi: 10.1080/1062936X.2019.1607899.

[4] D. P. P. Iv, “Studi Qsar Senyawa Turunan Triazolopiperazin Amida Sebagai Inhibitor Enzim Dipeptidil Peptidase-IV (DPP IV) Menggunakan Metode Semiempirik AM1,” *Bimipa*, vol. 23, no. 3, pp. 288–296, 2013.

[5] B. D. Green, P. R. Flatt, and C. J. Bailey, “Dipeptidyl peptidase IV (DPP IV) inhibitors: A newly emerging drug class for the treatment of type 2 diabetes,” *Diabetes and Vascular Disease Research*, vol. 3, no. 3, pp. 159–165, 2006, doi: 10.3132/dvdr.2006.024.

[6] I. M. Al-masri, M. K. Mohammad, and M. O. Taha, “Discovery of DPP IV inhibitors by pharmacophore modeling and QSAR analysis followed by in silico screening,” *ChemMedChem*, vol. 3, no. 11, pp. 1763–1779, 2008, doi: 10.1002/cmdc.200800213.

[7] P. Bharatam, D. Patel, L. Adane, A. Mittal, and S. Sundriyal, “Modeling and Informatics in Designing Anti-Diabetic Agents,” *Current Pharmaceutical Design*, vol. 13, no. 34, pp. 3518–3530, 2007, doi: 10.2174/138161207782794239.

[8] “Top 4 advantages and disadvantages of Support Vector Machine or SVM | by Dhiraj K | Medium.” https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107 (accessed Feb. 07, 2021).

[9] M. Lorca *et al.*, “Structure-activity relationships based on 3D-QSAR CoMFA/CoMSIA and design of aryloxypropanol-amine agonists with selectivity for the human β3-adrenergic receptor and anti-obesity and anti-diabetic profiles,” *Molecules*, vol. 23, no. 5, 2018, doi: 10.3390/molecules23051191.

[10] A. Allouche, “Software News and Updates Gabedit — A Graphical User Interface for Computational Chemistry Softwares,” *Journal of computational chemistry*, vol. 32, pp. 174–182, 2012, doi: 10.1002/jcc.

[11] U. Maulik and S. Bandyopadhyay, “Genetic algorithm-based clustering technique,” *Pattern Recognition*, vol. 33, no. 9, pp. 1455–1465, 2000, doi: 10.1016/S0031-3203(99)00137-5.

[12] S. Forrest, “Genetic algorithms: Principles of natural selection applied to computation,” *Science*, vol. 261, no. 5123, pp. 872–878, 1993, doi: 10.1126/science.8346439.

[13] C. Shen, L. Wang, and Q. Li, "Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method," *Journal of Materials Processing Technology*, vol. 183, no. 2–3, pp. 412–418, 2007, doi: 10.1016/j.jmatprotec.2006.10.036.

[14] J. Benesty, J. Chen, and Y. Huang, "On the importance of the pearson correlation coefficient in noise reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 757–765, 2008, doi: 10.1109/TASL.2008.919072.

[15] J. Shapiro, "Genetic algorithms in machine learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2049 LNAI, pp. 146–168, 2001, doi: 10.1007/3-540-44673-7_7.

[16] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the SMO algorithm for SVM regression," *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1188–1193, 2000, doi: 10.1109/72.870050.

[17] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "A fast iterative nearest point algorithm for support vector machine classifier design," *IEEE Transactions on Neural Networks*, vol. 11, no. 1, pp. 124–136, 2000, doi: 10.1109/72.822516.

[18] S. H. Min, J. Lee, and I. Han, "Hybrid genetic algorithms and support vector machines for bankruptcy prediction," *Expert Systems with Applications*, vol. 31, no. 3, pp. 652–660, 2006, doi: 10.1016/j.eswa.2005.09.070.

[19] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[20] J. T. Informasi, D. Kurniawan, P. Pascasarjana, M. Teknik, I. Universitas, and D. Nuswantoro, "Optimasi Algoritma Support Vector Machine (Svm)," vol. 9, no. April, pp. 38–49, 2013.

[21] A. Kulkarni, D. Chong, and F. A. Batarseh, *Foundations of data imbalance and solutions for a data democracy*. Elsevier Inc., 2020.

[22] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.