**Studi QSAR pada *In House Molecule* sebagai Inhibitor Papain-like Protease (PLpro) dari SARS-CoV-2 dengan menggunakan Metode Principal Component Analysis-Support Vector Machine**

**Tugas Akhir**

**diajukan untuk memenuhi salah satu syarat**

**memperoleh gelar sarjana**

**dari Program Studi Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**1301160357**

**Muhammad Alfi Al Ikhlas**

****

**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**2021**

LEMBAR PENGESAHAN

**Studi QSAR pada *In House* *Molecule* sebagai Inhibitor Papain-like Protease (PLpro) dari SARS-CoV-2 dengan menggunakan Metode Principal Component Analysis-Support Vector Machine**

**QSAR Study on In House Molecules as Papain-like Protease (PLpro) Inhibitors of SARS-CoV-2 by using Principal Component Analysis-Support Vector Machine Methods**

**NIM : 1301160357**

**Muhammad Alfi Al Ikhlas**

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

Universitas Telkom

Bandung, 24 Agustus 2021

Menyetujui

|  |  |  |
| --- | --- | --- |
| Pembimbing I, |  | Pembimbing II, |
|  |  |  |
| Isman Kurniawan, S.Pd., M.Si., M.Sc., Ph.D.  15870066 |  | Dra. Indiwiarti, M.Si.  98690022 |

|  |
| --- |
| Ketua Program Studi Sarjana Informatika,    Dr. Erwin Budi Setiawan, S.Si., M.T.  NIP: 00760045 |

LEMBAR PERNYATAAN

Dengan ini saya, Muhammd Alfi Al Ikhlas, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul *Studi QSAR pada In House Molecule sebagai Inhibitor Papain-like Protease (PLpro) dari SARS-CoV-2 dengan menggunakan Metode Principal Component Analysis-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, 24 Agustus 2021

Yang Menyatakan



Muhammad Alfi Al Ikhlas

Studi QSAR pada *In House Molecule* sebagai Inhibitor Papain-like Protease (PLpro) dari SARS-CoV-2 dengan menggunakan Metode Principal Component Analysis-Support Vector Machine

Muhammad Alfi Al Ikhlas, Isman Kurniawan, Indiwiarti

Fakultas Informatika, Universitas Telkom, Bandung

alfikhlas@student.telkomuniversity.ac.id, [ismanrkn@telkomuniversity.ac.id](mailto:ismanrkn@telkomuniversity.ac.id), indwiarti@telkomuniversity.ac.id

#### Abstrak

**Pandemi COVID-19 yang disebabkan oleh virus SARS-CoV-2 mempengaruhi seluruh dunia. Banyak sektor yang terganggu oleh pandemi seperti ekonomi, pendidikan, kesehatan, dan lain lain. Solusi yang diberikan saat ini, seperti *work from home* dan *study from home,* tidak bisa dilakukan pada semua sektor. Solusi terbaik saat ini hanyalah vaksin, yang belum bisa didapatkan oleh seluruh orang. Banyak faktor yang mempengaruhi penyebaran vaksin, mulai dari permasalahan biaya, waktu, dan stok vaksin, hingga beberapa orang dengan penyakit bawaan atau memiliki kondisi yang dianjurkan untuk menghindari vaksin. Oleh karena itu, penemuan obat yang setidaknya bisa meringankan gejala COVID-19 diperlukan untuk mengatasi permasalahan tersebut. Pencarian obat dilakukan dengan mencari molekul dengan aktivitas biologis terhadap PLpro yang merupakan protease penting dari replikasi virus. Proses desain obat bisa diakselerasi dengan mengimplementasikan metode machine learning pada model QSAR. Pemodelan dilakukan dengan menggunakan Principal Component Analysis untuk mereduksi fitur dan Support Vector Machine untuk mempelajari aktivitas biologis dari *in-house molecule* sebagai inhibitor PLpro. Pada penelitian ini ditemukan bahwa RBF merupakan kernel SVM paling optimal pada data tak dikenal dengan akurasi hingga 74% serta f-1 score yang menyentuh 72%.**

**Kata kunci : obat COVID-19, QSAR, Principal Component Analysis, Support Vector Machine**

# **Abstract**

**The world is currently facing COVID-19 pandemic which is caused by SARS-CoV-2. There are a lot of sectors that is affected by the current situation, namely economy, education, health, etc. The current solution, like *work from home* and *study from home*, cannot be applied to all sectors. The current best solution is vaccines, which is not entirely accessible to everyone. Things like cost, time, and vaccines stock affect the distribution of the vaccines, and for some people with health conditions or certain conditions, vaccines simply cannot be given to them. Finding a medicine would be an alternative way to solve those problems. A medicine candidate can be found by finding molecules which has a biological activity towards PLpro which is an essential protease for the virus replication. The process of designin medicine can be accelerated by implementing machine learning on QSAR model. The process of modelling are done by using Principal Component Analysis to reduce features and Support Vector Machine to study the biological activity of the *in-house molecule* as PLpro inhibitors. In this study, it is found that RBF is the optimal SVM kernel for unfamiliar data with an accuracy up to 74% and an f1-score of 72%.**

**Keywords: COVID-19 medicine, QSAR, Principal Component Analysis, Support Vector Machine**

#### Pendahuluan

Dunia saat ini sedang menghadapi pandemi COVID-19 yang disebabkan oleh virus SARS-CoV-2. Kondisi yang sama juga dialami oleh masyarakat Indonesia. Per tanggal 18 Maret 2021 total kasus COVID-19 di Indonesia sudah mencapai 1.495.614 kasus dengan angka kematian mencapai 46.905 jiwa [1]. Angka kematian yang hanya 3% mungkin tidak terlihat fatal, namun yang membuat COVID-19 berbahaya adalah tingkat penyebarannya yang tinggi. COVID-19 bisa dengan mudah menyebar dengan kontak langsung maupun secara tidak langsung [2]. Selain itu, virus ini pun bisa bertahan di udara hingga 3 jam [2]. Penyakit ini juga terhitung penyakit baru sehingga tidak ada orang yang imun terhadap penyakit ini. Selain itu, beberapa penderita COVID-19 mungkin hanya akan mengalami gejala ringan, namun pada beberapa orang virus ini bisa berakibat fatal. Virus ini lebih memungkinkan berakibat fatal pada orang tua atau seseorang yang memiliki kondisi medis lain seperti diabetes atau penyakit jantung.

Dengan mudahnya penyebaran virus COVID-19, tentu saja langkah pertama untuk menghadapi penyakit ini adalah memperlambat penyebaran COVID-19. Metode seperti *social distancing* serta *stay at home* sudah digunakan di berbagai negara [3], termasuk Indonesia. Namun metode ini bukan berarti metode terbaik dalam menghadapi penyakit COVID-19. Berbagai sektor seperti ekonomi serta pendidikan mengalami dampak yang cukup signifikan setelah adanya *work from home/study from home*. Solusi saat ini dari permasalahan tersebut adalah vaksin. Namun seseorang yang sudah mendapatkan vaksin belum tentu imun dari COVID-19. Selain itu, penyebaran vaksin juga memiliki kendala waktu dikarenakan masalah biaya beserta stok yang ada. Beberapa orang juga tidak bisa menerima vaksin dikarenakan memiliki kondisi khusus seperti wanita sedang hamil atau menyusui, penderita penyakit jantung, penderita autoimun, dll. Secara teori, jika semua orang yang bisa divaksin sudah mendapatkan vaksin, maka orang-orang dengan kondisi tersebut juga akan aman dari COVID-19. Kondisi ini biasa disebut dengan *herd immunity*. Namun untuk mencapai kondisi tersebut memerlukan waktu yang cukup lama. Oleh karena itu solusi yang paling mutakhir dari pandemi ini adalah dengan menemukan obat yang setidaknya bisa meringankan gejala COVID-19. Harapannya dengan ini penanganan kasus COVID-19 bisa lebih efektif dan cepat serta menurunkan tingkat kematian korban COVID-19.

COVID-19 mungkin terdengar baru di telinga masyarakat, namun pada kenyataannya COVID-19 merupakan coronavirus ketiga yang teridentifikasi menginfeksi manusia. Pada tahun 2002 hingga 2004 teridentifikasi penyakit *severe acute respiratory syndrome* (SARS) yang disebabkan *severe acute respiratory syndrome coronavirus* (SARS-CoV atau SARS CoV-1). Penyakit ini pertama kali ditemukan di Foshan, Guangdong, China, pada 16 November 2002 [4]. Terdapat 8,422 orang yang terinfeksi dengan *fatality rate* sekitar 10% [5]. Sedangkan pada tahun 2012 terdapat *Middle East respiratory syndrome* (MERS), yang bisa dikenal dengan flu unta [6]. Seperti namanya, penyakit ini pertama kali teridentifikasi di Timur Tengah, lebih tepatnya di Jeddah, Arab Saudi, pada tahun 2012. Penyakit ini menginfeksi setidaknya 1,700 orang dengan *fatality rate* sekitar 36% [7].

SARS-CoV-2 merupakan virus yang memiliki bentuk selubung beruntai tunggal dengan RNA *positive sense*. Protein spike pada SARS-CoV-2 mengatur masuknya virus ini kepada host yang dituju. Ketika virus ini masuk ke tubuh sebuah host, terdapat dua polyprotein yang ditranslasi, yaitu pp1a dan pp1ab. Kedua polyprotein ini oleh dihancurkan oleh dua protease milik virus SARS-CoV-2, yang pertama adalah *3 C-like protease* (3CLpro) dan yang kedua adalah *papain-like protease* (PLpro). Kedua protease ini merupakan bagian penting dalam proses replikasi virus SARS-CoV-2 pada tubuh host, oleh karena itu, kedua protease ini bisa digunakan sebagai target obat [7].

Tingginya tekanan dari segala sektor mulai dari kesehatan, pendidikan, bahkan ekonomi, menandakan bahwa pencarian obat COVID-19 perlu dipercepat. Metode QSAR merupakan metode yang cocok untuk mengakselerasi penelitian obat COVID-19. QSAR sendiri merupakan metode yang saat ini sering dipakai sebagai metode awal sebelum melakukan penelitian lebih jauh terhadap suatu penyakit [8]. Dengan memanfaatkan data molekul sintesis dari lab atau biasa dikenal *in-house molecule*, QSAR akan melakukan *Virtual Screening* (VS) untuk mengetahui relasi antara struktur molekul dengan aktivitas biologis [9]. Hal ini akan sangat membantu menghemat waktu, biaya, sumber daya, beserta tenaga dalam meneliti obat [8]. Beberapa penelitian sudah dilakukan untuk mencari kandidat obat COVID-19 dengan menggunakan metode QSAR seperti pada [7] dan [10]–[13].

Penelitian ini bertujuan untuk membuat model QSAR menggunakan machine learning. Proses machine learning dilakukan dengan menggunakan Principal Component Analysis (PCA) sebagai metode unuk mereduksi fitur serta Support Vector Machine sebagai algoritma pembelajaran aktivitas biologis *in-house molecule* sebagai inhibitor PLpro SARS-Cov-1.

#### Studi Terkait

##### Principal Component Analysis

Dataset yang berukuran besar sangat umum akhir-akhir ini, baik dari segi jumlah maupun dimensi. Hal ini mengakibatkan data tidak selalu mudah untuk dibaca. Principal Component Analysis (PCA) merupakan teknik untuk mereduksi dimensi dari sebuah dataset agar dataset tersebut lebih mudah dibaca namun secara bersamaan mencoba meminimalisir hilangnya informasi yang penting. PCA melakukan hal tersebut dengan cara membuat variabel baru yang tidak berkorelasi namum berhasil memaksimalkan variansi. PCA bisa membuat variabel tersebut dengan cara memecahkan permasalahan eigenvalue/eigenvector dan membuat variabel baru berdasarkan data yang ada, dengan ini PCA merupakan metode yang efektif dan adaptif.

|  |  |
| --- | --- |
| . | () |

Langkah pertama dalam Principal Component Analysis adalah menemukan Principal Component (PC) yang bisa dilakukan dengan melakukan langkah pada rumus (1) [14] . Setelah itu, masing-masing PC akan diurutkan berdasarkan nilai eigenvalue dan eigenvector dari yang terbesar dan terkecil. Jika jumlah fitur yang ingin digunakan sudah ditentukan sebelumnya, PCA hanya akan menyimpan PC dengan jumlah yang sesuai dengan jumlah fitur tersebut.

##### Support Vector Machine

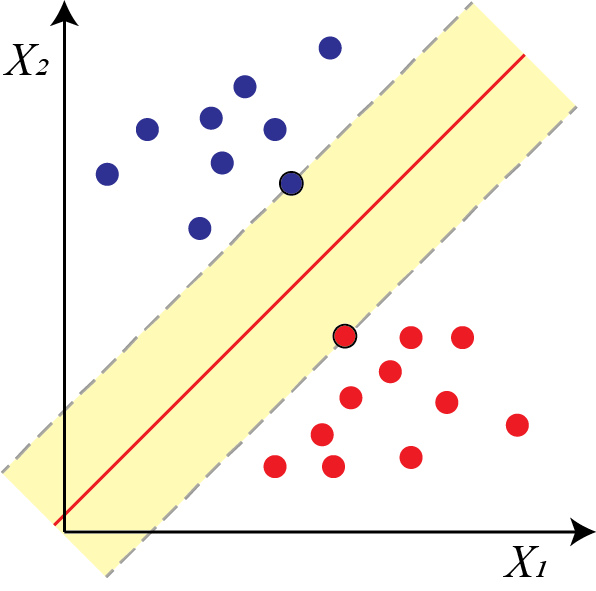
Support Vector Machine (SVM) merupakan algoritima machine learning yang belajar dari suatu contoh atau *dataset* untuk melabeli objek [15]. Sebagai contoh, SVM bisa dilatih untuk mendeteksi penipuan kartu kredit dengan cara menganalisis ratusan hingga ribuan data aktivitas penggunaan kartu kredit yang terhitung penipuan dan non-penipuan, lalu mengkategorikan masing-masing aktivitas ke dalam kelompoknya masing-masing. SVM juga sudah banyak diaplikasikan pada beragam bidang biologi. Penggunaan umum SVM dalam ranah biomedis adalah klasifikasi otomatis dalam profil ekspresi gen microarray. Secara teori, SVM bisa mempelajari profil ekspresi yang diturunkan dari sampel tumor atau dari cairan perifer lalu mendapatkan diagnosis atau prognosis berdasarkan data tersebut [16]. Penggunaan lain SVM pada bidang biologi juga mencakup objek yang beragam seperti protein, DNA, ekspresi microarray, dan spektrometri massa [17].

Jika dijelaskan secara sederhana, SVM merupakan algoritma yang mencari *hyperplane* terbaik sebagai pemisah antara dua buah kelas data. *Hyperplane* terbaik ditemukan dengan cara mengukur *margin* dari *hyperplane* yang ada lalu mencari jarak paling maksimal. *Margin* sendiri merupakan jarak antara *hyperplane* dengan titik terdekat dari masing-masing kelas. Titik paling dekat tersebut juga biasa disebut sebagai *support vector* . Sebuah hyperplane bisa ditulis sebagai sebuah set titik yang memenuhi Rumus (2).

|  |  |
| --- | --- |
|  | () |

Dimana adalah vektor normal dari *hyperplane.* Parameter menentukan *offset hyperplane* dari posisi aslinya.

SVM sendiri memiliki performa yang cukup menjanjikan dalam memodelkan QSAR berdasarkan beberapa penelitian seperti [18],[19], dan [20]. Pada penelitian protease inhibitor HIV di [18], SVM memiliki hasil yang lebih superior dibandingkan dengan Multiple Linear Regression (MLR) dan Artificial Neural Network (ANN). Hasil dari penelitian [19] juga membuktikan bahwa SVM memiliki prospek yang cukup bagus dalam QSAR modelling.



*Gambar 1 Hyperplane SVM*

#### Sistem yang Dibangun

Secara garis besar, program bisa dibagi menjadi preprocessing dataset, pembuatan model prediksi, dan validasi model. Garis besar dari program bisa dilihat pada gambar 2.

A picture containing diagram

Description automatically generated

*Gambar 2 Alur pembuatan program*

##### Dataset

Data yang digunakan pada studi ini didapatkan dari [7]. Data tersebut terdiri dari beragam senyawa yang memliki aktivitas penghambatan terhadap PLpro SARS-CoV. Senyawa yang tidak memiliki aktivitas penghambatan tidak akan digunakan pada studi ini. Selain itu, duplikat dari molekul yang sudah ada juga tidak akan digunakan pada studi ini. Selain itu, ditentukan juga *‘Activity treshold’* sebagai dasar perhitungan. *‘Activity treshold’* sendiri merupakan rata-rata dari nilai SARS-Cov PLpro. Senyawa yang memiliki nilai kurang dari *‘activity treshold’* akan dilabeli sebagai tidak aktif atau *Inactive*, sedangkan senyawa yang memiliki nilai lebih dari *‘activity treshold’* akan dianggap sebagai menjanjikan sehingga dilabeli aktif atau *Active*. Berdasarkan persyaratan tersebut, dari 90 molekul didapatkan 40 senyawa yang dilabeli aktif dan 50 senyawa yang dilabeli tidak aktif.

Dikarenakan data yang didapatkan dari [7] hanya berbentuk notasi SMILES dan nilai aktivitas, data tersebut harus melalui *pre-processing* terlebih dahulu. Notasi SMILES yang didapatkan dari data lalu diproses oleh *mordred descriptor* untuk mendapakan fitur data. Setelah itu, data dibagi menjadi data test dan data train dengan rasio 3:7. Setelah itu, data juga melalu proses scaling menggunakan *Robust Scaler*. Dengan ini data sudah siap untuk melalui proses selanjutnya.

##### Seleksi Fitur

Hasil dari *pre-processing* sebelumnya membuat data menjadi lebih rapih, namun dengan jumlah fitur sebanyak 1329 fitur, proses *machine learning* tidak akan optimal dikarenakan jumlah dimensi yang terlalu banyak. Oleh karena itu diperlukan proses seleksi fitur untuk mengurangi dimensi dari data tersebut. Proses seleksi fitur dilakukan dengan menggunakan Principal Component Analysis (PCA). Dari PCA, dibuatlah enam model data. Enam model data tersebut masing-masing memilik jumlah fitur sebanyak 5, 10, 15, 20, 25, dan 30 fitur. Hal ini dikarenakan setelah 30 fitur hampir seluruh data sudah bisa diwakili. Penambahan fitur lebih dari 30 hanya akan membuat model data menjadi redundan. Setelah itu, masing-masing model data ini akan digunakan pada masing-masing model prediksi lalu dihitung akurasinya. Model data dengan akurasi tertinggi untuk satu model prediksi akan digunakan sebagai model data untuk model prediksi tersebut.

##### Model Prediksi

Model prediksi yang digunakan untuk memprediksi nilai aktivitas dibuat dengan menggunakan Support Vector Machine (SVM). Ada tiga kernel SVM yang digunakan sebagai model pada studi ini, yaitu *RBF, Linear,* dan *Polynomial*. Setelah itu, data yang sebelumnya sudah direkduksi dimensinya menggunakan PCA lalu digunakan kepada tiga model prediksi tersebut lalu dihitung akurasinya. Setelah didapatkan model data yang paling optimal untuk masing-masing model prediksi, proses selanjutnya adalah *hyperparameter tuning.*

Pada *hyperparameter tuning*, setiap model prediksi melalui proses *Grid Search* yang dimana setiap model prediksi tersebut mengalami perubahan parameter guna menemukan parameter yang paling optimal untuk masing-masing model tersebut. Terdapat tiga parameter yang mengalami proses tuning, yaitu C, Gamma, dan Degree. Pada kasus C, dipilih enam nilai eksponensial kelipatan 10. Angka ini dipilih demi mencakup banyak area dalam proses *grid search* tanpa perlu terlalu banyak meningkatkan kompleksitas pencarian. Pada kasus Gamma, hanya ada dua kemungkinan sehingga keduanya digunakan pada *grid search*. Setelah itu, untuk kasus Degree digunakan empat pangkat dengan pangkat maksimal 5. Detail dari parameter yang digunakan beserta cakupannya bisa dilihat pada tabel 1.

*Tabel 1 Hyperparameter Tuning*

|  |  |  |  |
| --- | --- | --- | --- |
| *Parameter* | *RBF* | *Linear* | *Polynomial* |
| C | [0.001, 0.01, 0.1, 1, 10, 100] | [0.001, 0.01, 0.1, 1, 10, 100] | [0.001, 0.01, 0.1, 1, 10, 100] |
| Gamma | ['scale', 'auto'] | - | ['scale', 'auto'] |
| Degree | - | - | [2, 3, 4, 5] |

##### Validasi

Untuk mengetahui seberapa akurat model yang sudah dibuat maka proses validasi perlu dilakukan. Proses validasi dilakukan dengan cara mengaplikasikan model yang sudah dilatih menggunakan data train kepada data test yang merupakan data baru bagi model tersebut. Ada empat parameter yang digunakan sebagai alat ukur proses validasi, yaitu *Accuracy, Precision, Recall,* dan *F-1 Score*. Bagaimana proses perhitungan dari keempat parameter tersebut bisa dilihat pada rumus (3), (4), (5) dan (6). Semua proses perhitungan validasi menggunakan konsep *Confusion Matrix* yang bisa dilihat pada Tabel 2.

*Tabel 2 Confusion Matrix*

|  |  |  |
| --- | --- | --- |
|  | Prediksi | |
| Aktual | TRUE | FALSE |
| TRUE | True Positive (TP) | False Positive (FP) |
| FALSE | False Negative (FN) | True Negative (TN) |

|  |  |
| --- | --- |
|  | (3) |
|  | (4) |
|  | (5) |
|  | (6) |

#### Hasil dan Pembahasan

##### Seleksi Fitur

Pada proses seleksi fitur, enam model data diaplikasikan kepada tiga model prediksi untuk dibandingkan hasilnya. Masing-masing model prediksi memiliki hasil yang berbeda terkait model data terbaik. Untuk model *RBF*, model data dengan 5 fitur merupakan model terbaik, sedangkan pada model *Linear*, model data dengan 15 fitur merupakan yang terbaik, dan terakhir pada model *Polynomial*, model data dengan 30 fitur merupakan model data terbaik. Detail dari hasil seleksi fitur bisa dilihat pada Tabel 3 dan grafik pada Gambar 3.

*Chart, line chart

Description automatically generated*

*Gambar 3* *Grafik Perbandingan Komponen PCA*

*Tabel 3 Komponen PCA Terbaik*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Kernel | Best PCA Component | Accuracy | Average ± std |
| 1 | RBF | 5 | 63% | 59% ± 3.9% |
| 2 | Linear | 15 | 66% | 60% ± 4% |
| 3 | Polynomial | 30 | 58% | 57% ± 1.2% |

##### Hyperparameter Tuning

Hampir dari semua model prediksi memiliki perubahan pada parameter yang dimilikinya. Semua model prediksi mengalami perubahan pada parameter C yang bernilai *default* 1. Pada RBF, parameter C yang paling optimal adalah 10 sedangkan pada Linear dan Polynomial, 100 merupakan parameter C paling optimal. Setelah itu pada kasus Gamma, Polynomial memliki nilai parameter *Auto* sebagai hasil yang lebih optimal dibandingkan nilai *default*-nya yang merupakan *Scale* sedangkan RBF tetap menggunakan parameter *default*. Untuk parameter Degree, tidak ada perubahan sama sekali pada semua model prediksi. Hasil dari *Hyperparameter Tuning* yang dilakukan bisa dilihat pada Tabel 4.

*Tabel 4 Hasil Hyperparameter Tuning*

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | RBF | Linear | Polynomial |
| C | 10 (1) | 100 (1) | 100 (1) |
| Gamma | Scale (Scale) | - | Auto (Scale) |
| Degree | - | - | 3 (3) |
| Keterangan : nilai yang berada di dalam tanda kurung ( ) merupakan nilai *default* | | | |

Perbandingan model sebelum *tuning* dan sesudah *tuning* bisa dilihat pada Gambar 4. Perbandingan model dilakukan pada kedua data test dan juga data train. Pada kasus data test, model RBF tidak mengalaim perubahan yang signifikan, sedangkan pada model Linear, terdapat penurunan akurasi sekitar ± 10%. Berbeda dengan model Linear, model Polynomial mengalami penaikan akurasi sebanyak ± 4%.

Untuk kasus data train, akurasi model Polynomial mengalami penaikan akurasi hingga 100% dari model sebelumnya yang hanya memiliki akurasi sekitar 68%. Sedangkan pada model Linear, model setelah *tuning* mengalami penurunan akurasi sebanyak ± 1%. Yang terakhir, model RBF bisa menyentuh akurasi hingga 90% setelah mengalami proses *tuning*, terdapat kenaikan sekitar ± 10% dari model sebelumnya yang memiliki akurasi sekitar 80%.

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart, bar chart  Description automatically generated |
| *Gambar 4 Perbandingan Tuned dan Untuned pada Data Test (kiri) dan Data Train (kanan)* | |

##### Validasi Model

Pada proses validasi, didapatkan bahwa model RBF memiliki hasil yang terhitung stabil dibandingkan model Linear dan model Polynomial*.* Model Polynomial memiliki akurasi yang sangat tinggi bisa mencapai 100% pada data train, namun ketika dihadapi oleh dataset baru, akurasi yang didapatkan menurun sangat drastis hingga akurasi yang didapatkan hanyalah 55%. Sekalipun model Linear tidak memiliki tingkat akurasi setinggi model Polynomial, model Linear juga mengalami permasalahan dengan model Polynomial. Ketika ditemukan oleh dataset baru, model Linear memiliki akurasi 37% dari yang sebelumnya 79% ketika dihadapkan oleh data train. Tabel 5 memberikan detail lebih lanjut terkait hasil validasi.

Tabel 5 Rangkuman Hasil Validasi

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Model* | *Kernel* | *TN* | *FP* | *FN* | *TP* | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
| *Traning set* | | | | | | | | | |
| 1 | RBF | 31 | 6 | **0** | **26** | 90.48% | 81.25% | 100% | 89.66% |
| 2 | Linear | 33 | 4 | 9 | 17 | 79.36% | 80.95% | 65.38% | 72.34% |
| 3 | Polynomial | **37** | **0** | **0** | **26** | **100%** | **100%** | **100%** | **100%** |
| *Test set* | | | | | | | | | |
| 1 | RBF | **11** | **2** | **5** | **9** | **74.07%** | **81.81%** | **64.28%** | **72%** |
| 2 | Linear | 6 | 7 | 10 | 4 | 37.04% | 36.36% | 28.57% | 32% |
| 3 | Polynomial | 10 | 3 | 9 | 5 | 55.56% | 62.5% | 35.71% | 45.45% |

Pada proses tuning, parameter yang berubah hanyalah nilai dari parameter C dan Gamma. Pada parameter C, nilai tersebut digunakan untuk menentukan *decision boundary*. Semakin besar nilai C, semakin besar *penalty* untuk setiap adanya titik data yang termisklasifikasi, begitu juga sebaliknya. Sedangkan Gamma mempengaruhi nilai yang dipengaruhi oleh satu titik data. Nilai Gamma yang rendah mengakibatkan klusterisasi yang luas, hal ini mengakibatkan banyak data yang akan terklusterisasi dalam satu kluster. Sedangkan nilai Gamma yang besar membutuhkan titik data dengan jarak yang sangat dekat satu sama lain untuk bisa dimasukan kedalam satu kluster, hal ini sering mengakibatkan overfit. Setelah tuning, kernel RBF memiliki nilai C=10 yang lebih tinggi dari nilai *default* C=1, hal ini mengakibatkan *decision boundary* yang dibuat bisa menerima error lebih sedikit dibanding parameter normal. Pada kasus kernel Linear, Nilai C=100 justru mengakibatkan akurasi yang didapatkan menurun dari parameter normal. Akurasi yang didapatkan saat dihadapi data test juga sangat buruk. Bisa disimpulkan bahwa pada kasus kernel Linear, data tidak bisa dipishakan secara linear dan *decision boundary* yang dibuat terlalu tegas sehingga gagal membagi data dengan baik. Pada kasus Polynomial, pemisahan data saat training kemungkinan besar mengalami *overfit* dengan nilai parameter C=100 dan Gamma=Auto sehingga ketika dipertemukan dengan data point baru pada data test, perfroma yang sebelumnya mencapai 100% menurun hingga 55%. Selain itu, RBF dengan bentuk kernelnya yang menyerupai distribusi normal kemungkinan memiliki bentuk yang menyerupai data sehingga memiliki akurasi yang superior dibanding kedua kernel lainnya.

#### Kesimpulan

Terdapat beberapa hasil yang bisa disimpulkan berdasarkan penelitian yang sudah dilakukan. Pertama, Principal Component Analysis berhasil mereduksi jumlah fitur dengan menyimpan informasi sebanyak mungkin. Jumlah fitur yang awalnya mencapai 1329 fitur berhasil direduksi hingga maksimal fitur yang ada hanyalah 30 fitur. Selain itu, Support Vector Machine juga memiliki hasil yang cukup menjanjikan dalam memodelkan QSAR dengan kernel RBF sebagai kernel dengan performa terbaik dalam pengaplikasiannya. Ketika dipertemukan dengan model data training, akurasi yang didapatkan sangat tinggi hingga mencapai 90%. Selain itu ketika dipertemukan dengan model data test, akurasi yang dimiliki masih terbilang akurat dengan akurasi hingga 74% dan f-1 score hingga 72%. Berdasarkan hasil tersebut, bisa disimpulkan bahwa RBF memiliki kemampuan prediksi yang cukup tinggi untuk data point yang familiar pada data train, namun cukup fleksibel untuk tetap bisa memprediksi data point asing pada data test. Dengan ini bisa disimpulkan bahwa RBF merupakan kernel yang cocok untuk memodelkan QSAR untuk pencarian obat COVID-19 dengan performanya yang stabil.

#### Daftar Pustaka

[1] Tim Kawal COVID-19, “Kawal informasi seputar COVID-19 secara tepat dan akurat.,” *Kawalcovid19.Id*, 2020. https://kawalcovid19.id/ (accessed Mar. 18, 2021).

[2] M. R. Tfi, M. R. Hamblin, and N. Rezaei, “Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ’ s public news and information,” *Clin. Chim. Acta*, vol. 508, no. January, pp. 254–266, 2020, [Online]. Available: www.elsevier.com/locate/cca Review.

[3] K. G. Emerson, “Coping with being cooped up: Social distancing during COVID-19 among 60+ in the United States,” *Rev. Panam. Salud Publica/Pan Am. J. Public Heal.*, vol. 44, pp. 1–7, 2020, doi: 10.26633/RPSP.2020.81.

[4] C. Law, “A Chronicle on the SARS Epidemic,” *Chinese Law Gov.*, vol. 36, no. 4, pp. 12–15, 2003, doi: 10.2753/clg0009-4609360412.

[5] WHO, “Update 49 - SARS case fatality ratio, incubation period,” *WHO*, 2003. https://www.who.int/csr/sars/archive/2003\_05\_07a/en/ (accessed May 21, 2021).

[6] Richard Lloyd Parry, “Travel alert after eighth camel flu death,” *The Times UK*, 2015. https://www.thetimes.co.uk/article/travel-alert-after-eighth-camel-flu-death-2k8j83mzgq2 (accessed May 21, 2021).

[7] S. A. Amin, K. Ghosh, S. Gayen, and T. Jha, “Chemical-informatics approach to COVID-19 drug discovery: Monte Carlo based QSAR, virtual screening and molecular docking study of some in-house molecules as papain-like protease (PLpro) inhibitors,” *J. Biomol. Struct. Dyn.*, vol. 0, no. 0, pp. 1–10, 2020, doi: 10.1080/07391102.2020.1780946.

[8] B. J. Neves, R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade, “QSAR-based virtual screening: Advances and applications in drug discovery,” *Front. Pharmacol.*, vol. 9, no. NOV, pp. 1–7, 2018, doi: 10.3389/fphar.2018.01275.

[9] S. Kwon, H. Bae, J. Jo, and S. Yoon, “Comprehensive ensemble in QSAR prediction for drug discovery,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019, doi: 10.1186/s12859-019-3135-4.

[10] N. Jha *et al.*, “Deep Learning Approach for Discovery of In Silico Drugs for Combating COVID-19,” *J. Healthc. Eng.*, vol. 2021, pp. 1–13, 2021, doi: 10.1155/2021/6668985.

[11] J. Ivanov *et al.*, “Quantitative structure−activity relationship machine learning models and their applications for identifying viral 3Clpro- And RDRP-targeting compounds as potential therapeutics for Covid-19 and related viral infections,” *ACS Omega*, vol. 5, no. 42, pp. 27344–27358, 2020, doi: 10.1021/acsomega.0c03682.

[12] X. Yu, “Prediction of inhibitory constants of compounds against SARS-CoV 3CLpro enzyme with 2D-QSAR model,” *J. Saudi Chem. Soc.*, vol. 25, no. 7, p. 101262, Jul. 2021, doi: 10.1016/J.JSCS.2021.101262.

[13] A. A. Ishola, O. Adedirin, T. Joshi, and S. Chandra, “QSAR modeling and pharmacoinformatics of SARS coronavirus 3C-like protease inhibitors,” *Comput. Biol. Med.*, vol. 134, p. 104483, Jul. 2021, doi: 10.1016/J.COMPBIOMED.2021.104483.

[14] I. Djakaria, S. Guritno, and S. H. Kartiko, “Visualisasi Data Iris Menggunakan Analisis Komponen Utama dan Analisis Komponen Utama Kernel Visualization of Iris Data Using Principal Component Analysis and Kernel Principal Component Analysis,” *J. ILMU DASAR*, vol. 11(1), no. 1, pp. 31–38, 2010.

[15] B. E. Boser, V. N. Vapnik, and I. M. Guyon, “Training Algorithm Margin for Optimal Classifiers,” *Perception*, pp. 144–152, 1992.

[16] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006, doi: 10.1038/nbt1206-1565.

[17] W. S. Noble, “Support vector machine applications in computational biology,” 2004, pp. 71–92.

[18] R. Darnag, B. Minaoui, and M. Fakir, “QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression,” *Arab. J. Chem.*, vol. 10, pp. S600–S608, 2017, doi: https://doi.org/10.1016/j.arabjc.2012.10.021.

[19] H. Mei, Y. Zhou, G. Liang, and Z. Li, “Support vector machine applied in QSAR modelling,” *Chinese Sci. Bull.*, vol. 50, no. 20, pp. 2291–2296, 2005, doi: 10.1007/BF03183737.

[20] S. M. Marunnan *et al.*, “Development of MLR and SVM Aided QSAR Models to Identify Common SAR of GABA Uptake Herbal Inhibitors used in the Treatment of Schizophrenia,” *Curr. Neuropharmacol.*, vol. 15, no. 8, pp. 1085–1092, Nov. 2017, doi: 10.2174/1567201814666161205131745.

##### Lampiran

Lampiran dapat berupa detil data dan contoh lebih lengkapnya, data-data pendukung, detail hasil pengujian, analisis hasil pengujian, detail hasil survey, surat pernyataan dari tempat studi kasus, screenshot tampilan sistem, hasil kuesioner dan lain-lain.