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, 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

1. 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.

2. 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.

$$\begin{aligned} PC_1 &= a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p \\ PC_2 &= a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p \\ &\vdots \\ PC_k &= a_{1k}X_1 + a_{2k}X_2 + \cdots + a_{pk}X_p \end{aligned} \tag{1}$$