TESIS

PENINGKATAN KINERJA ALGORITMA *CORONA VIRUS DISEASE OPTIMISATION* MENGGUNAKAN *COMPETITION STRATEGY* UNTUK SELEKSI FITUR BERBAGAI DIMENSI DATA

Performance Improvement of Corona Virus Disease Optimisation Algorithm Using Competition Strategy for Feature Selection of Various Data Dimensions

MURISNAN
D082212002

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA
DEPARTEMEN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

PENGAJUAN TESIS

PENINGKATAN KINERJA ALGORITMA *CORONA VIRUS DISEASE OPTIMISATION* MENGGUNAKAN *COMPETITION STRATEGY* UNTUK SELEKSI FITUR BERBAGAI DIMENSI DATA

Tesis
Sebagai Salah Satu Syarat untuk Mencapai Gelar Magister
Program Studi Magister Teknik Informatika

Disusun dan diajukan oleh

MURISNAN
D082212002

Kepada

FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2024

TESIS

PENINGKATAN KINERJA ALGORITMA *CORONA VIRUS DISEASE OPTIMISATION* MENGGUNAKAN *COMPETITION STRATEGY* UNTUK SELEKSI FITUR BERBAGAI DIMENSI DATA

MURISNAN
D082212002

Telah dipertahankan di hadapan Panitia Ujian Magister Pada Tanggal 10 Juli 2024
dan dinyatakan telah memenuhi syarat kelulusan

Pada

Program Studi Teknik Informatika
Departemen Teknik Informatika
Fakultas Teknik
Universitas Hasanuddin
Gowa

Mengesahkan:

PembimbingUtama,

Dr. Amil Ahmad Ilham, S.T., M.IT.
NIP. 197310101998021001

Pembimbing Pendamping,

Adnan, S.T., M.T. Ph.D
NIP. 197404262005011002

Ketua Program Studi
Magister Teknik Informatika,

Dr. Ir. Zahir Zainuddin, M.Sc.
NIP. 196404271989101002

Dekan Fakultas Teknik
Universitas Hasanuddin,

Prof. Dr.Eng.Ir.Muhammad Isran Ramli, M.T.,IPM.,ASEAN.Eng.
NIP. 197309262000121002

PERNYATAAN KEASLIAN TESIS
DAN PELIMPAHAN HAK CIPTA

Yang bertanda tangan di bawah ini

Nama : Murisnan

Nomor Mahasiswa : D082212002

Program Studi : Magister Teknik Informatika

Dengan ini menyatakan bahwa, tesis berjudul "**PENINGKATAN KINERJA ALGORITMA *CORONA VIRUS DISEASE OPTIMISATION* MENGGUNAKAN *COMPETITION STRATEGY* UNTUK SELEKSI FITUR BERBAGAI DIMENSI DATA**" adalah benar karya Saya dengan arahan dari komisi pembimbing (Dr. Amil Ahmad Ilham, S.T.,M.IT. sebagai Pembimbing Utama dan Adnan, S.T., M.T. Ph.D sebagai Pembimbing Pendamping). Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi manapun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan dari Penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka tesis ini. Sebagian dari isi tesis ini telah dipublikasikan di Prosiding *2023 IEEE International Conference on Communication, Networks and Satellite (*COMNETSAT), Halaman 359-364, dan DOI : 10.1109/COMNETSAT59769.2023.10420710 sebagai artikel dengan judul "*Optimal Feature Selection Using Modified COVID Optimization Algorithm*".

Dengan ini Saya melimpahkan hak cipta dari karya tulis Saya berupa tesis ini kepada Universitas Hasanuddin.

Gowa, 10 Juli 2024

Yang menyatakan

Murisnan

KATA PENGANTAR

Bismillah, Assalamu'alaikum Warohmatullahi Wabarokatuh, Alhamdulillah Puji dan Syukur Penulis panjatkan atas Kehadirat Allah Subhanallahu Wa Ta'ala karena berkat Rahmat dan Karunia-Nya sehingga Penulis dapat menyelesaikan Karya Ilmiah berupa Tesis ini yang berjudul "**Peningkatan Kinerja Algoritma *Corona Virus Disease Optimisation* Menggunakan *Competition Strategy* Untuk Seleksi Fitur Berbagai Dimensi Data**". Penyusunan Tesis ini merupakan salah satu syarat untuk memperoleh gelar Magister Komputer (M.Kom) pada Program Studi Magister Teknik Informatika, Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin.

Penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya atas dukungan dan semangat yang diberikan serta membantu Penulis baik secara langsung ataupun tidak langsung dalam menyelesaikan Karya Ilmiah ini. Penulis ingin mengucapkan terima kasih kepada :

1 Allah SWT karena berkat Rahmat dan Karunia-Nya Penulis berhasil menyusun Tesis ini dengan baik.
2 Keluarga Penulis, Istri dan Tiga orang Putri cantik yang masih kecil saat Penulis menempuh Pendidikan Magister, terima kasih atas doa dan bantuan moral dan materi selama Penulis belajar di Departemen Teknik Informatika Universitas Hasanuddin.
3 Direktur dan Staff Unit Kepegawaian Politeknik Negeri Ujung Pandang yang telah memberikan dukungan dan izin belajar di Program Studi Magister Teknik Informatika.
4 Bapak Dr. Ir. Zahir Zainuddin, M.Sc. sebagai Kordinator Prodi Magister Teknik Informatika yang selalu memberikan arahan kepada Penulis agar cepat menyelesaikan segala rangkaian proses akademik di Departemen Teknik Informatika.
5 Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. dan Adnan, S.T., M.T..Ph.D. Selaku Pembimbing I dan Pembimbing II yang telah memberikan ilmu dan

arahan serta dukungan yang penuh sehingga memudahkan dalam penyusunan dan penulisan tesis ini.

6 Seluruh Bapak dan Ibu Dosen di Departemen Teknik Informatika yang telah memberikan ilmu selama Penulis kuliah di Program Studi Magister Teknik Informatika.

7 Seluruh Staff Administrasi Program Studi Magister Teknik Informatika yang telah memberikan bantuan administrasi selama kuliah di Fakultas Teknik, Universitas Hasanuddin.

8 Pengurus dan Anggota Himpunan Mahasiswa Magister Teknik Informatika yang senantiasa memberikan dukungan dan semangat kepada Penulis.

9 Serta seluruh teman-teman Angkatan 3, 4 dan 5 Mahasiswa Program Studi Magister Teknik Informatika yang sangat luar biasa dan hebat- hebat, serta patut dibanggakan.

Penulis mohon maaf yang sebesar-besarnya apabila terdapat kekurangan dalam penyusunan dan penulisan Tesis ini. Kritik dan Saran yang membangun Penulis harapkan dari para Pembaca dan Peneliti serta Akademisi untuk perbaikan dan pembelajaran dikemudian hari, semoga tulisan ini dapat memberikan manfaat yang besar dan seluas-luasnya untuk kemajuan Ilmu pengetahuan di Indonesia pada umumnya dan ilmu dibidang komputer pada khususnya. Terima kasih, Wasalamu ‘Alaikum Warohmatullahi Wabarokatuh.

Gowa, 10 Juli 2024

Penulis

ABSTRAK

MURISNAN. *Peningkatan Kinerja Algoritma Corona Virus Disease Optimisation Menggunakan Competition Strategy Untuk Seleksi Fitur Berbagai Dimensi Data.* (Dibimbing oleh **Amil Ahmad Ilham dan Adnan**).

Berbagai dimensi data yang bervariasi menjadi tantangan terbesar saat ini disebabkan oleh perkembangan dan bertambahnya data yang begitu cepat setiap harinya. Selain itu keinginan untuk mendapatkan informasi dari tumpukan data ini juga menjadi suatu hal yang sangat penting. Untuk mendapatkan informasi yang diinginkan dari tumpukan data berbagai dimensi ini sehingga perlu dilakukan penggalian data atau data mining. Salah satu bagian dari teknik data mining adalah pra-pemrosesan data yaitu seleksi fitur. Seleksi fitur merupakan proses untuk memilih fitur-fitur yang penting dalam suatu dataset yaitu dengan mengurangi fitur-fitur yang tidak sesuai, tidak relevan atau tidak perlu yang dianggap akan memperburuk kinerja dalam proses klasifikasi data dan waktu pemrosesannya. Dengan memodifikasi Algoritma COVID Optimization untuk memilih subset fitur yang relevan dari berbagai jenis dimensi data menggunakan konsep *Competition Strategy* yang disematkan pada Algoritma optimasi yaitu *Coronavirus Disease Optimization Algorithm* untuk menangani masalah seleksi fitur yang relevan. Melalui konsep *Competition Strategy* ini yang bertujuan untuk meningkatkan pencarian global dan mampu memberikan kinerja untuk menghasilkan subset fitur yang menjanjikan dan mencapai akurasi klasifikasi yang lebih baik. Kinerja Algoritma yang diusulkan diuji pada enam dataset dari *UCI Machine Learning Data Repository* dipilih berdasarkan tingkat akurasi yang rendah pada penelitian sebelumnya. Hasil eksperimen menunjukkan bahwa kinerja Algoritma yang diusulkan lebih unggul dari Algoritma versi Biner COVID Optimization dalam menangani seleksi fitur yang relevan sehingga mencapai tingkat akurasi yang sangat baik.

Kata Kunci *: Feature Selection, Wrapper Aproach, Metaheuristic Algorithm, Optimization Algorithm, Evolutionary Algorithm.*

ABSTRACT

MURISNAN. *Performance Improvement of Corona Virus Disease Optimisation Algorithm Using Competition Strategy for Feature Selection of Various Data Dimensions.* (*Supervised by* **Amil Ahmad Ilham and Adnan**).

The biggest challenge today is dealing with various data dimensions due to their rapid development and daily increase. Furthermore, the desire to obtain information from this pile of data is critical. To get the desired information from this pile of data of various dimensions, data extraction or data mining is required. Data pre-processing, or feature selection, is one part of the data mining technique. Feature selection is the process of selecting important features in a dataset by reducing inappropriate, irrelevant, or unnecessary features that are considered to worsen performance in the data classification process and processing time. To solve the problem of choosing the right features, change the COVID Optimisation Algorithm to pick a subset of relevant features from different types of data dimensions. This can be done by using the Competition Strategy idea that is built into the optimisation algorithm, which is called the Coronavirus Disease Optimisation Algorithm. This competition strategy concept aims to improve global search performance by producing promising feature subsets and classification accuracy. We tested the proposed algorithm's performance on six datasets from the UCI Machine Learning Data Repository, which we selected due to their low accuracy rates in previous studies. The experimental results show that the proposed algorithm outperforms the Binary COVID Optimization version in handling the selection of relevant features, achieving excellent accuracy.

Keywords: *Feature Selection, Wrapper Aproach, Metaheuristic Algorithm, Optimisation Algorithm, Evolutionary Algorithm*.

DAFTAR ISI

Halaman

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR SINGKATAN, ISTILAH & SIMBOL

Lambang/Singkatan		Arti dan Keterangan
ACO	=	*Ant Colony Optimisation*
COVIDOA	=	*Corona Virus Disease Optimisation Algorithm*
PSO	=	*Particle Swarm Optimisation*
SMOTE	=	*Synthetic Minority Oversampling Techniue*
ENN	=	*Edited Nearest Neighbor*
KNN	=	*K- Nearest Neighbor*
RGA	=	*Rival Genetic Algorithm*
BCOVIDOA	=	*Binary Covid Optimisation Algorithm*
COVIDOA-C	=	*Covid Optimisation Algorithm-Competition*
O	=	Big Oh

BAB I
PENDAHULUAN

1.1 Latar Belakang

Era Revolusi Industri 4.0 telah banyak menghasilkan data berdimensi tinggi yang menjadi sumber penting bagi banyak penelitian terkait klasifikasi, pengelompokan, pengenalan pola dan prediksi (Elgamal *et al.*, 2020). Sejumlah data yang dihasilkan setiap harinya adalah data yang berupa teks, angka, audio, video, dan gambar. Data ini berasal dari berbagai bidang seperti kesehatan, pertanian, transportasi, keuangan, pendidikan dan bidang lainnya (Abiodun *et al.*, 2021).

Peningkatan volume dan dimensi data menyebabkan banyak masalah, seperti nilai yang hilang, data yang tidak seimbang, dan fitur-fitur data yang tidak relevan (Elgamal *et al.*, 2020). Kumpulan data yang tidak seimbang dan berisi fitur-fitur yang tidak relevan atau berlebihan, sangat tidak bermanfaat bagi model prediksi atau model klasifikasi, hal ini juga akan meningkatkan waktu komputasi dan mengurangi kinerja klasifikasi bagi model pendekatan yang menggunakan machine learning atau data mining (Rong, Gong and Gao, 2019).

Pendekatan *machine learning* dan *data mining* yang efektif menjadi perhatian yang lebih bagi para Peneliti saat ini diberbagai disiplin ilmu seperti kedokteran, bioinformatika, penambangan teks, pengolahan citra, atau pengenalan wajah (Khurma *et al.*, 2022). Data mining merupakan pendekatan penting dalam proses penemuan pengetahuan dan membantu dalam mengekstrak pola, model, dan aturan dari dataset. Pendekatan ini terdiri dari seleksi data, pembersihan data, integrasi data, reduksi data, penambangan data, evaluasi pola dan representasi pengetahuan yang baru. (Christo *et al.*, 2022). Pada tahapan pra-pemrosesan antara lain meliputi bagian pembersihan data, integrasi data, transformasi data, dan reduksi data.

Sebelum melakukan klasifikasi, tahap pra-pemrosesan data adalah hal yang harus dilakukan untuk memastikan bahwa dataset yang kita miliki memiliki kualitas yang baik (Lan *et al.*, 2018). Meskipun hasil dari tugas analisis data tetap

tergantung pada beberapa faktor seperti seleksi fitur, seleksi algoritma, teknik pengambilan sampel, dan lain-lain (Khan and Hoque, 2020).

Pada tahap pra-pemrosesan data, yang sangat penting diperhatikan adalah pada fase seleksi fitur atau dengan istilah yang lain yaitu seleksi fitur pada dataset. Proses seleksi fitur adalah mencari fitur yang berkorelasi dan menghapus fitur yang berlebihan atau tidak berkorelasi dari sebuah kumpulan fitur. Seleksi fitur juga dikenal sebagai seleksi variabel, seleksi fitur, atau seleksi subset variabel. Diberbagai penelitian metode seleksi fitur berkinerja baik dalam menyederhanakan model dan membantu waktu komputasi pelatihan lebih efisien (Rong, Gong and Gao, 2019).

Beberapa metode seleksi fitur telah dikembangkan untuk mendapatkan subset fitur terbaik. Salah satu pengembangan metode seleksi fitur yang paling populer saat ini yaitu seleksi fitur yang menggunakan Algoritma Metaheuristik (Agrawal *et al.*, 2021). Namun Secara umum metode seleksi fitur diklasifikasikan ke dalam tiga kategori yaitu metode *filter, wrapper dan embedded* (Agrawal *et al.*, 2021). Berdasakan penelitian Agrawal dkk (2021), metode *wrapper* memberikan hasil yang lebih baik dibandingkan dengan metode *filter*, seleksi fitur dengan Algoritma Metaheuristik termasuk dalam kategori metode *wrapper*.

Algoritma Metaheuristik adalah metode optimasi yang mendapatkan solusi optimal atau mendekati optimal dari masalah optimasi. Algoritma metaheuristik fleksibel dan lugas karena konsepnya yang sederhana dan implementasinya yang mudah. Algoritmanya dapat dimodifikasi dengan mudah sesuai dengan masalah tertentu (Agrawal *et al.*, 2021).

Berdasarkan perilakunya, Algoritma Metaheuristik dibagi menjadi empat kategori yaitu algoritma berbasis evolusi, algoritma berbasis *swarm intelligence*, algoritma berbasis fisika, dan algoritma berbasis *human behavior* (Agrawal *et al.*, 2021).

Algoritma berbasis evolusi ini terinspirasi dari evolusi alami dan memulai prosesnya dengan populasi solusi yang dihasilkan secara acak, contoh Algoritma yang populer dalam kategori ini yaitu Algoritma Genetika (AG), Tabu Search (TS), Evolusi Diferensial (ED) (Agrawal *et al.*, 2021).

Kemudian Algoritma yang berbasis *swarm intelligence* ini terinsiprasi oleh perilaku sosial hewan seperti serangga, lebah, ikan, burung, serigala, contoh Algoritmanya antara lain Particle Swarm Optimization (PSO), Artificial Bee Colony Optimization (ABC), Ant Colony Optimization (ACO), Bat Algorithm (BA), Gravitational Search Algorithm (GSA), Firefly Algorithm (FA) (Abiodun *et al.*, 2021).

Selanjutnya Algoritma berbasis fisika terinspirasi oleh aturan Fisika, contohnya Harmoni Search Algorithm (HSA) dan Simulated Annealing (SA) (Agrawal *et al.*, 2021), dan yang terakhir yaitu Algoritma berbasis human behavior yang terinspirasi oleh perilaku manusia, contohnya Teaching learning-based optimization algorithm (TLBO) dan League Championship algorithm (LCA) (Agrawal *et al.*, 2021).

Algoritma optimasi terbaru metaheuristik saat ini yang menangani seleksi fitur dengan pendekatan biner yaitu *Binary Coronavirus Disease Optimization Algorithm* (BCOVIDOA) oleh (Khalid *et al.*, 2022) yang merupakan modifikasi dari Algoritma optimasi *Coronavirus Disease Optimization Algorithm* (COVIDOA) dari hasil penelitian (Khalid, Hosny and Mirjalili, 2022).

Algoritma COVIDOA adalah algoritma optimasi evolusioner yang terinspirasi dari mekanisme replikasi pertikel virus Corona saat menyerang tubuh manusia. COVIDOA telah terbukti memiliki kinerja yang baik jika dibandingkan dengan algoritma metaheuristik terkenal lainnya dengan kemampuan eksplorasi dan ekploitasinya yang tinggi (Khalid, Hosny and Mirjalili, 2022).

Namun teorema *No Free Lunch* (NFL) yang dikemukakan oleh (Lockett, 2020) secara logis telah membuktikan dan mengklaim bahwa tidak ada algoritma metaheuristik yang paling cocok untuk menyelesaikan semua jenis masalah pengoptimalan. Dengan kata lain ada beberapa Algoritma yang menyelesaikan masalah dengan baik, tetapi ada juga beberapa Algoritma yang memberikan kinerja buruk (Hosseini *et al.*, 2021).

Teorema NFL tersebut dapat dijadikan landasan bahwa dari penelitian sebelumnya yang telah mengusulkan Algoritma COVIDOA versi biner dengan sebutan BCOVIDOA telah ditemukan ada beberapa dataset yang tingkat

akurasinya rendah, oleh sebab itu dapat dikatakan bahwa algoritma BCOVIDOA belum berkinerja secara optimal. Hal tersebut diduga bahwa Algoritma COVIDOA versi biner belum optimal kinerjanya saat melakukan seleksi fitur sehingga hasil akurasi klasifikasinya rendah pada beberapa dataset.

Dengan memperhatikan potensi pengembangan yang ada pada Algoritma optimasi COVIDOA yang tergolong masih baru ini sehingga perlu dilakukan penelitian lebih lanjut untuk mencoba memodifikasi dengan cara menggabungkan, menerapkan atau mengintegrasikan metode/algoritma lain yang mampu memberikan perubahan kinerja terhadap Algoritma COVIDOA ini.

Oleh karenanya pada penelitian ini mencoba bereksperimen menggunakan metode atau konsep *Competition Strategy* yang akan diintegrasikan ke dalam Algoritma COVIDOA untuk menangani masalah seleksi fitur.

Konsep *Competition Strategy* yang digunakan oleh Too dan Abdullah (2021) dalam penelitiannya mengungkapkan bahwa penerapan *Competition Strategy* tidak hanya dapat meningkatkan kinerja sistem tetapi juga mengurangi kompleksitas komputasi.

1.2 Rumusan Permasalahan

Berdasarkan uraian latar belakang penelitian ini, sehingga dirumuskan suatu masalah yaitu Bagaimana cara meningkatkan kinerja Algoritma COVIDOA untuk menangani seleksi fitur sehingga mampu meningkatkan hasil akurasi klasifikasi berbagai dimensi dataset dari penelitian sebelumnya ?

1.3 Tujuan Penelitian

Berdasarkan deskripsi latar belakang dan rumusan masalah maka tujuan penelitian ini adalah akan mengusulkan Algoritma COVIDOA yang dimodifikasi menggunakan konsep *Competition Strategy* untuk melakukan seleksi fitur pada suatu dataset, sehingga akan terbentuk Model klasifikasi dengan dugaan sementara bahwa model yang diusulkan mampu memberikan kinerja yang optimal dalam tugas menangani seleksi fitur dan melakukan proses klasifikasi yang lebih baik.

1.4 Manfaat Penelitian

Penelitian yang akan dilakukan tentunya diharapkan memberikan manfaat pengetahuan kepada Penulis secara langsung, kemudian diharapkan mampu memberikan informasi pengetahuan baru dan berharga bagi para Pembaca dan Peneliti lain bahwa metode yang diusulkan untuk menangani seleksi fitur mampu meningkatkan kinerja klasifikasi dalam bidang ilmu data mining dan machine learning.

1.5 Ruang Lingkup

Agar Penelitian ini menjadi lebih terarah dan fokus serta memberikan kontribusi pengetahuan yang lebih spesifik, maka perlu diuraikan beberapa ruang lingkup atau batasan masalah penelitian antara lain:

- Dataset yang digunakan untuk klasifikasi sama dengan dataset yang digunakan pada penelitian COVIDOA versi biner (Khalid *et al.*, 2022) yang diperoleh dari repository UCI Machine Learning. Yang merupakan dataset dari variasi ukuran dimensi.
- Algoritma klasifikasi yang digunakan sama dengan Algoritma klasifikasi yang digunakan pada penelitian sebelumnya yaitu KNN.
- Tools dan kode pemrograman yang akan digunakan untuk seleksi fitur dan klasifikasi adalah MATLAB R2022a.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Masalah seleksi fitur adalah salah satu tugas paling menantang dalam pembelajaran mesin (Agrawal *et al.*, 2021). *Exhautive search, greedy search, random search* dan berbagai metode lainnya telah diusulkan. Agrawal *et al*. (2021) mengemukakan bahwa Sebagian besar metode mengalami konvergensi prematur, kompleksitas yang sangat besar serta biaya komputasi yang tinggi.

Sejak tahun 1970-an banyak upaya telah dilakukan untuk mengevaluasi metode seleksi fitur, seleksi fitur dibagi menjadi 4 kelompok yaitu *filter*, *wrapper*, *hybrid* dan *embedded* (Rostami, Berahmand and Forouzandeh, 2021).

Agrawal *et al.* (2021) juga memaparkan bahwa metode hybrid menggabungkan *filter* dan *wrapper*. Metode *wrapper* memberikan hasil yang lebih baik dibandingkan dengan metode *filter*, tetapi metode *wrapper* dalam prosesnya lambat atau dengan kata lain memerlukan waktu komputasi yang tinggi. Untuk menjawab masalah komputasi yang tinggi dalam seleksi fitur, Algoritma metaheuristik hadir untuk memberikan solusi dengan metode optimasi yang mendapatkan solusi optimal (mendekati optimal). Perilaku algoritma metaheuristik adalah stochastic yaitu dimulai dengan proses optimasi dengan menghasilkan solusi secara random atau acak.

Algoritma metaheuristik terbaru tahun 2022 versi Biner COVIDOA yang diusulkan oleh Asma Khalid dkk (Khalid *et al.*, 2022) yang menangani seleksi fitur diuji dan dikomparasikan dengan beberapa algoritma seleksi fitur berbasis metaheusristik lainnya seperti GA, PSO, DE, WOA, WOASA, GWOPSO, HH, GWO, dan AOA (Khalid *et al.*, 2022). Setelah dilakukan pengujian dengan menggunakan 26 jenis dataset hasil akurasi rata-rata mencapai 92%, Best fitness 0,0898, Rata-rata fitness 0.0920, Standar Deviasi 0,0019, dan ukuran seleksi fitur rata-rata 147,15. Hasil yang diperoleh mengungkapkan efisiensi dari algoritma yang diusulkan serta menunjukkan kemampuan eksplorasi dan ekploitasi yang kuat. Hasil uji komputasi tidak tampak dieksplor oleh Penulis pada tulisannya.

Penelitian terkait lainnya yang dilakukan oleh Jingwei dkk (Too and Abdullah, 2021) menggunakan Algoritma RGA & FRGA diuji pada kumpulan dataset publik yang diperoleh dari *UCI Machine Learning Repositori* sebanyak 20 dataset, dan 3 dataset *dari Arizona State University*. Dari 23 dataset yang digunakan untuk melakukan validasi kinerja seleksi fitur terdiri dari 3 kategori jenis dataset, yaitu kategori dimensi rendah (jumlah fitur < 30), kategori dimensi menengah (jumlah fitur lebih besar atau sama dengan 30 dan kurang dari 400), dan kategori dimensi tinggi (jumlah fitur > 400). Hasil akurasi dari Algoritma ini (Too and Abdullah, 2021) menyaingi Algoritma metaheuristik populer lainnya dengan hasil rata-rata 95%. Mengenai waktu komputasi Algoritma FRGA-T memberikan waktu komputasi singkat dengan waktu terendah 0.0396 detik pada salah satu dataset. Jika berdasarkan rata-rata waktu komputasi dari 23 dataset maka waktu komputasinya 64,46 detik.

Rostami dkk (Rostami, Berahmand and Forouzandeh, 2021) mengusulkan algoritma varian baru dari Genetic Algorithm (GA) dengan sebutan *Community Detection Based Genetic Algorithm for Feature Selection* (CDGAFS) untuk seleksi fitur berbagai dimensi data. Algoritma CDGAFS yang diusulkan ini menggunakan klasifikasi KNN, SVM, dan AdaBoost kemudian dibandingkan dengan algoritma populer lainnya seperti PSO, ABC, dan ACO. CDGAFS berada peringkat pertama dengan hasil akurasi rata-rata dari klasifikasi 89,89 % dari hasil uji menggunakan 6 dataset berbagai dimensi data. Mengenai hasil eksekusi waktu komputasi CDGAFS dengan nilai rata-rata 37.20 detik yang merupakan rata-rata tercepat dibandingkan PSO, ABC, dan ACO.

Penelitian lainnya terkait seleksi fitur yang menggunakan pendekatan hybrid yang dilakukan (Al-Wajih *et al.*, 2021) mengusulkan *Binary Grey Wolf Optimizer* (BGWO) dengan *Harris Hawks Optimizer* (HHO) dengan sebutan HBGWOHHO dan menggunakan 18 dataset dengan hasil akurasi dan waktu komputasi yang sangat baik, rata-rata akurasi 92 % dan waktu komputasinya paling rendah 2,54 detik.

Kemudian (Bhattacharyya *et al.*, 2020) juga mengusulkan algoritma hybrid baru metaheuristik yaitu MA-HS yang merupakan kombinasi dari *Mayfly*

Algorithm (MA) dan *Harmony Search* (HS). Algoritma yang diusulkan untuk menangani seleksi fitur dengan menggunakan 18 dataset dari UCI repository dan membandingkan dengan 12 Algoritma populer yang menangani masalah fitur seleksi. Hasil ekperimen dari algoritma yang diusulkan akurasi 100% pada 8 dataset namun tidak menjelaskan hasil komputasi pada penelitian tersebut.

Penelitian lainnya dari Xueting Cui dkk (Cui *et al.*, 2020) tahun 2020 mengusulkan algoritma hybrid dengan sebutan *Hybrid Imroved Dragonfly Algorithm* (HIDA). Algoritma ini menggabungkan *Dragonfly Algorithm* (DA) dan *Maximum Relevance and Minimum Redundancy* (MRMR). Kinerja Algoritma HIDA divalidasi pada 10 dataset gene expression dan 8 dataset dari UCI machine learning repositori. Klasifikasi yang digunakan adalah SVM. Hasil penelitian (Cui *et al.*, 2020) Algoritma HIDA unggul diantara 7 algoritma lainnya dalam hal akurasi pada salah satu dataset, akurasi memberikan nilai 100 %. Masalah waktu komputasi tidak dibahas dalam penelitian ini.

Berikutnya penelitian dari El- Sayed dkk (El-Kenawy and Eid, 2020) tahun 2020 mengusulkan metode hybrid untuk menangani seleksi fitur yaitu *Grey Wolf Optimization* dan *Particle Swarm Optimization* (GWOPSO). Penelitian ini menggunakan 17 dataset dari UCI machine learning repositori. Algoritma klasifikasi yang digunakan yaitu KNN. Namun hasil klasifikasinya tidak dibahas, hanya menampilkan hasil kesalahan klasifikasi terendah, best fitness, worst fitness, dan standar deviasi. Waktu pemrosesan (detik) dihasilkan dari penelitian ini GWOPSO unggul diberbagai dataset hasil terendah untuk untuk setiap dataset pada pengujian menggunakan dataset hepatitis dengan skor 2.8 detik.

Metode seleksi fitur hibrid lainnya yang diusulkan oleh Sedighe Abasabadi dkk (Abasabadi *et al.*, 2022) tahun 2022, yang disebut GArank&rand, yang menggabungkan seleksi fitur *filter* yang diusulkan (SLI-ÿ) dan pendekatan seleksi fitur berbasis GA pembungkus. Awalnya, SLI-ÿ digunakan untuk menghapus 99% fitur yang tidak relevan pada fase pertama. Kemudian, GA menggunakan fitur paling relevan yang dihitung oleh SLI-ÿ untuk mengoptimalkan solusi fase pertama. Penelitian ini menggunakan 11 kumpulan data terkenal (termasuk 7 kumpulan data

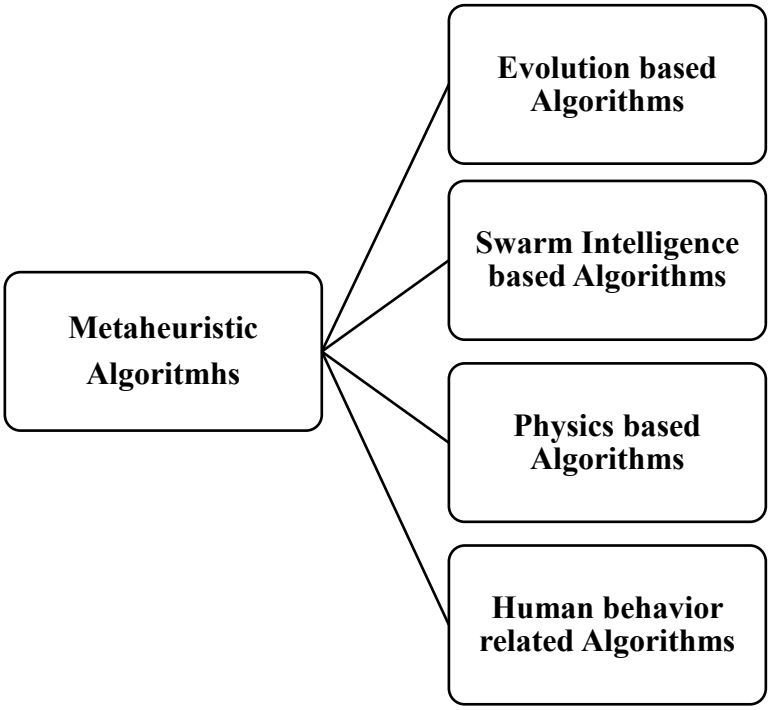
dimensi tinggi dan 4 kumpulan data standar) untuk evaluasi kriteria pengukuran. Hasil percobaan menunjukkan bahwa SLI-ÿ mampu menghasilkan hasil yang lebih baik daripada pemeringkat lainnya (yang dipertimbangkan dalam penelitian ini) pada semua dataset. Selain itu, SLI-ÿ memiliki dampak yang signifikan terhadap performa GA dalam hal akurasi klasifikasi dan jumlah fitur yang dipilih. Selain itu, waktu eksekusi GArank&rand menurun secara signifikan ketika 1% dari fitur peringkat terbaik dipilih untuk generasi populasi GA.

Dalam penelitian yang diusulkan Mohamed G. El-Shafiey dkk (El-Shafiey *et al.*, 2022) tahun 2022, yaitu Algoritma genetika (GA) dan Partikel Swarm Optimization (PSO) berbasis Random Forest (RF), yang disebut GAPSO-RF, dikembangkan dan digunakan untuk memilih fitur optimal yang dapat meningkatkan akurasi prediksi penyakit jantung dengan sebutan GAPSO-RF yang diusulkan mengimplementasikan analisis statistik multivariat pada langkah pertama untuk memilih fitur paling signifikan yang digunakan pada populasi awal. Setelah itu, strategi mutasi diskriminatif diterapkan di GA. GAPSO-RF menggabungkan GA yang dimodifikasi untuk penelusuran global dan PSO untuk penelusuran lokal. Metode yang diusulkan mencapai akurasi tinggi masing-masing 95,6% dan 91,4% pada dataset Cleveland dan Statlog. Setelah itu, hasil dari metode FS yang diusulkan dibandingkan dengan hasil tanpa menggunakan FS dan menemukan bahwa akurasinya lebih baik. Selain itu, ini mengungguli metode canggih yang ada pada kumpulan data yang sama. Selanjutnya, analisis komparatif dilakukan antara GAPSO-RF dan GA konvensional dan menemukan bahwa pendekatan yang diusulkan mengungguli GA konvensional. Kemudian algoritma klasifikasi yang digunakan yaitu Random Forest (RF). Oleh karena itu, hasil percobaan kami menegaskan bahwa pendekatan yang diusulkan meningkatkan proses pengambilan keputusan para praktisi diagnosis penyakit jantung. Namun untuk kompleksitas waktu metode ini bukanlah yang terbaik.

2.2 Algoritma Metaheuristik

Algoritma metaheuristik adalah metode optimasi yang mendapatkan solusi optimal (mendekati optimal) dari masalah optimasi. Algoritma ini adalah teknik turunan bebas dan memiliki kesederhanaan serta fleksibilitas. Perilaku algoritma metaheuristik adalah stokastik; mereka memulai proses optimasi mereka dengan menghasilkan solusi acak. Tidak perlu menghitung turunan dari ruang pencarian seperti pada teknik pencarian gradien. Algoritma metaheuristik fleksibel dan lugas karena konsepnya yang sederhana dan implementasinya yang mudah. Algoritma dapat dimodifikasi dengan mudah sesuai dengan masalah tertentu (Agrawal et al., 2021).

Algoritma Metaheuristik telah berhasil digunakan untuk mengatasi masalah klasifikasi. Metaheuristik diperkenalkan ke dalam seleksi fitur di berbagai bidang karena kemampuan dan kinerja pencarian globalnya yang luar biasa. Metaheuristik telah diterapkan untuk banyak tantangan optimasi dunia nyata, termasuk load balancing dalam jaringan telekomunikasi dan jadwal penerbangan, masalah pengiriman beban ekonomi, pemilihan gen dalam klasifikasi kanker dalam domain medis (Abiodun et al., 2021).



Gambar 1 Jenis Algoritma Metaheurstik

Evolution Based Algorithm, yaitu Algoritma berbasis evolusi ini terinspirasi dari evolusi alami dan memulai prosesnya dengan populasi solusi yang dihasilkan secara acak. Dalam jenis algoritma ini, solusi terbaik disatukan untuk menciptakan individu baru. Individu baru dibentuk dengan menggunakan mutasi, crossover dan memilih solusi terbaik. Yang paling yang populer dalam kategori ini adalah Algoritma Genetika (GA) yang didasarkan pada teknik evolusi Darwin. Ada algoritma lain seperti strategi evolusi, pemrograman genetik, pencarian tabu, evolusi diferensial (Agrawal *et al.*, 2021).

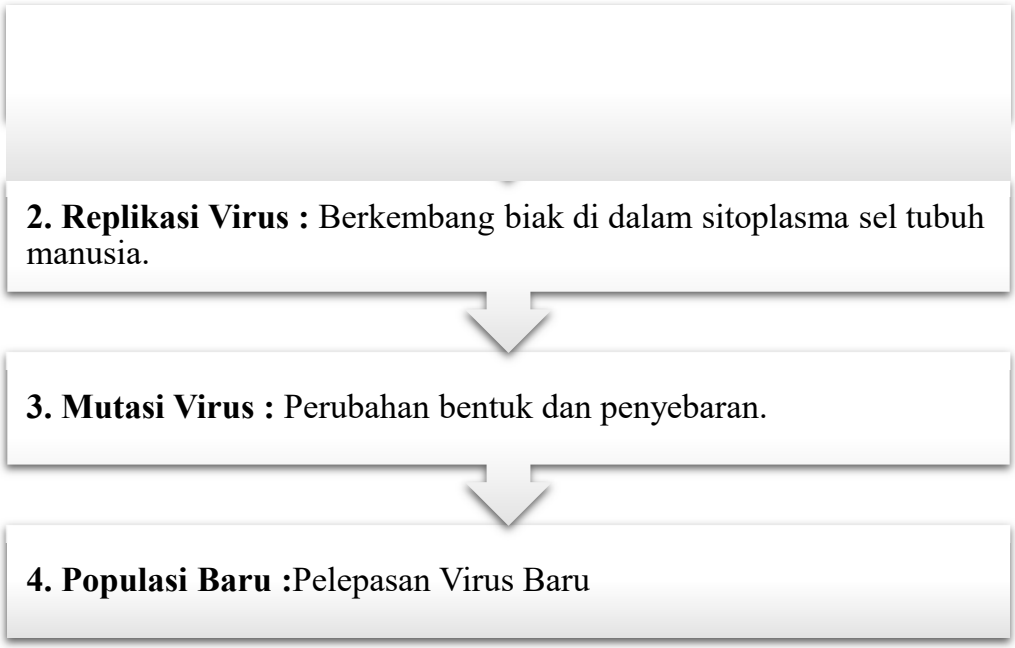
Swarm Intelligence Based Algorithm, adalah Algoritma berbasis kecerdasan kawanan, kumpulan, atau gerombolan. Algoritma ini terinspirasi oleh perilaku sosial serangga, hewan, ikan atau burung dll. Teknik yang populer adalah *Particle Swarm Optimization* (PSO) yang dikembangkan oleh Kennedy dan Eberhart. Hal ini terinspirasi dari perilaku sekelompok burung yang terbang melintasi ruang pencarian dan menemukan lokasi (posisi) terbaiknya. *Ant Colony Optimization* (ACO) yang terinspirasi dari koloni semut. *Honey Bee Swarm Optimization Algorithm* (HBSO) yang terisnpirasi dari kawanan lebah madu. Monkey Optimization yang terinspirasi dari kawanan monyet dll, merupakan contoh dari algoritma kecerdasan kawanan/gerombolan (Agrawal *et al.*, 2021).

Physics Based Algorithm, Algoritma berbasis fisika Ini terinspirasi oleh aturan fisika di alam semesta. *Simulated Annealing, Harmoni Search Algorithm* (HAS), *Gravitational Search Algorithm* (GSA), *Galaxy Based Search Algorithm* (GBSA) dll berada di bawah algoritma berbasis fisika (Agrawal *et al.*, 2021).

Human Behavior Related Algorithm, Algoritma terkait perilaku manusia, model ini murni terinspirasi oleh perilaku manusia. Setiap manusia memiliki caranya sendiri dalam melakukan aktivitas yang mempengaruhi kinerjanya. Algoritma yang populer adalah *Teaching learning-based optimization algorithm* (TLBO), *League Championship Algorithm, Human Inspired Algorithm* (HIA), *Social Emotional Optimization Algorithm* (SEOA) dll (Agrawal *et al.*, 2021).

2.3 Corona Virus Disease Optimisation Algorithm (COVIDOA)

COVIDOA merupakan Algoritma pengoptimalan berbasis evolusi yang dibuat oleh (Khalid, Hosny and Mirjalili, 2022). COVIDOA terinspirasi dari mekanisme replikasi partikel Virus Corona saat menyerang tubuh manusia. Penjelasan literatur dari penelitian sebelumnya yang telah mengusulkan COVIDOA dalam versi biner untuk menangani seleksi fitur bahwa ada empat tahapan dari Algoritma COVIDOA yaitu :



Gambar 2 Tahapan Algoritma COVIDOA

Khalid *et al*.(2022) mendeskripsikan model matematika dari COVIDOA sebagai berikut:

1. Inisialisasi populasi solusi diinisialisasi secara acak, dan dievaluasi untuk setiap solusi. Solusi tersebut kemudian diurutkan sesuai dengan fungsi fitness, dan solusi pertama dianggap sebagai solusi terbaik.
2. Tahap replikasi virus melalui teknik frameshifting Ribosomal, frameshifting adalah proses ketika kerangka pembacaan tertentu dari molekul RNA bergeser ke kerangka pembacaan lain untuk menyediakan

urutan protein baru. Frameshifting menghasilkan beberapa protein virus. Dalam algoritma COVIDOA, solusi (partikel virus) dipilih untuk replikasi. Teknik frameshifting menghasilkan beberapa protein virus yang kemudian digabungkan untuk membentuk partikel virus baru. Jenis frameshifting yang populer adalah frameshifting yang dimodelkan sebagai berikut :

- **+1 teknik framesifhting**

Nilai solusi Parent digeser ke arah kanan sebesar 1, dan nilai diposisi pertama ditetapkan sebagai nilai acak dalam rentang [minVal maxVal].

$$Sk\ (1) = rand\ (minVal,\ maxVal) \quad (1)$$

$$Sk\ (2{:}D) = P\ (1{:}D\ -1) \quad (2)$$

- **-1 teknik framesifhting**

Nilai solusi Parent digeser ke arah kanan sebesar 1, dan nilai diposisi pertama ditetapkan sebagai nilai acak dalam rentang [minVal maxVal].

$$Sk\ (D) = rand\ (minVal,\ maxVal) \quad (3)$$

$$Sk\ (1{:}D{-}1) = P\ (2{:}D) \quad (4)$$

Keterangan :

Dimana *minVal* dan *maxVal* adalah nilai minimal dan maximal untuk variabel di setiap solusi.

Sk = Nilai Protein ke-k

P = Nilai Parent Solusi

D = Dimensi Masalah (Jumlah variabel dalam setiap solusi)

3. Mutasi diterapkan pada solusi yang dibuat pada langkah sebelumnya (Langkah 2) untuk menghasilkan solusi mutasi baru. Persamaanya sebagai berikut.

$$Z(i) = \begin{cases} r, & if\ rand\ (0,1) < MR \\ X(i), & otherwise \end{cases} \quad (5)$$

Keterangan:

X = Solusi sebelum mutasi

Z = Solusi termutasi

X(i) dan Z(i) adalah elemen ke- i dalam solusi lama dan baru

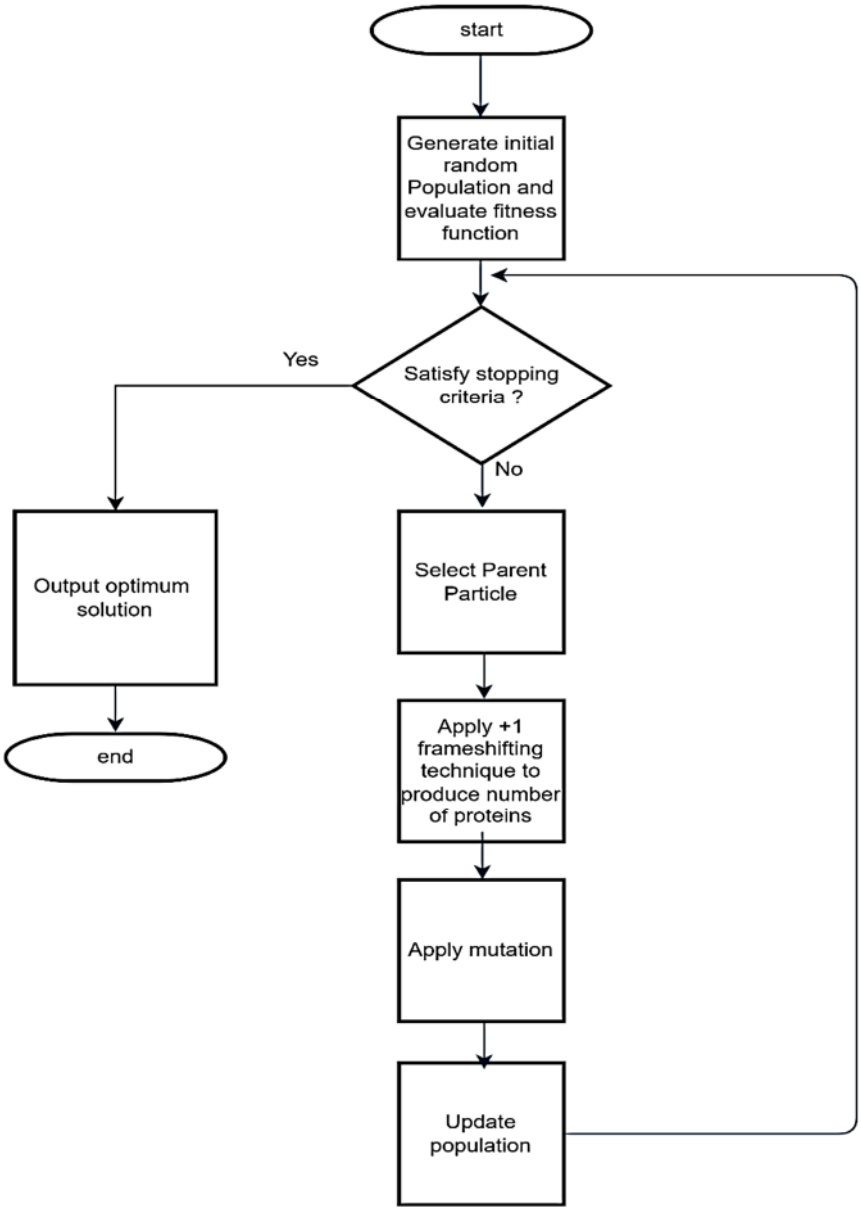
i = 1, ...,D,

r = Nilai Random dengan range *(minVal, maxVal)*

MR = Mutation Rate

4. Fungsi objektif dievaluasi untuk solusi baru dan Populasi diperbarui untuk generasi berikutnya (Solusi dengan nilai fitnes tertinggi digunakan, dan yang lain dihilangkan).
5. Ulangi langkah (2-4) untuk Populasi baru, hingga kriteria terminasi terpenuhi atau jumlah iterasi maksimum tercapai.
6. Hingga output Solusi terbaik ditemukan.

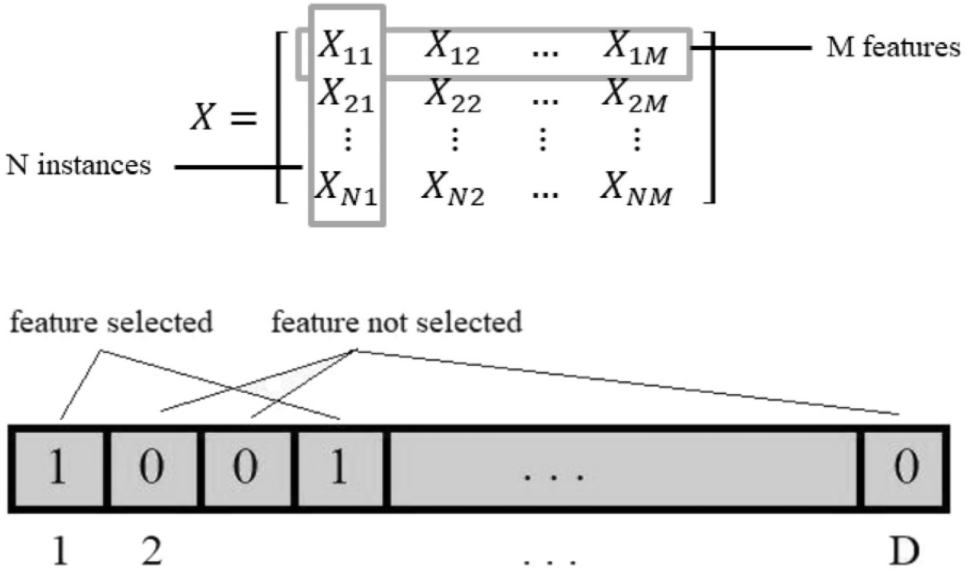
Adapun model Flowchart Algoritma COVIDOA sebagai berikut :



Gambar 3 Flowchart Algoritma COVIDOA

2.4 *Binary Corona Virus Disease Optimisation Algorithm* (BCOVIDOA)

Dalam Algoritma versi biner yang diusulkan oleh Khalid *et al.*, (2022) untuk menangani seleksi fitur, setiap solusi diwakili oleh vektor satu dimensi yang hanya berisi angka (0) dan angka (1), yang mana angka (1) menunjukkan bahwa fitur tersebut dipilih, lalu untuk angka (0) diabaikan atau fitur yang tidak terpilih. Representasi biner dari solusi COVID untuk dataset dengan jumlah fitur disimbolkan (*D*). Berikut representasi dari masalah seleksi fitur (Khalid *et al.*, 2022).



Gambar 4 Representasi Dataset dan Fitur Biner

Model Matematika BCOVIDOA sebagai berikut (Khalid *et al.*, 2022) :

1. Inisialisasi Populasi secara acak

$$xi = minVal_i + \alpha_i \ (maxVal_i - minVal_i) \ , i=1,2,.....D \quad (6)$$

Keterangan:

x_i = Solusi ke i, dalam Populasi

α_i = Nilai acak antara 0 dan 1

$maxVal_i$ - $minVal_i$ = batas atas dan batas bawah

(Untuk Versi Biner batas atas = 1, batas bawah =0)

Untuk BCOVIDOA, setiap nilai solusi (x_i) diubah menjadi representasi biner (Khalid *et al.*, 2022) menggunakan teknik binarisasi yaitu fungsi sigmoid bipolar berikut persamaannya:

$$S(x_i) = \frac{1}{1+e^{-x_i}} \quad (7)$$

Keterangan;

$S(x_i)$ = Solusi ke i, Populasi

$x_{binary} = 1$ if rand $\geq S(x_i)$, jika tidak = 0

2. Replikasi virus dalam versi biner diterapkan dalam persamaan sebagai berikut :

$$S_k(1) = \begin{cases} 1 & if\ rand\ (0,1) < 0.5 \\ 0 & otherwise \end{cases} \quad (8)$$

$$S_k(2:D) = P(1:D-1) \quad (9)$$

Keterangan:

S_k = Nilai Protein yang dihasilkan ke -k

rand = Nilai acak antara 0 dan 1

P = Nilai Parent Solusi

D = Dimensi Masalah (Jumlah variabel dalam setiap solusi)

3. Mutasi dalam versi biner diterapkan dalam persamaan sebagai berikut:

$$Z_i = \begin{cases} \begin{cases} 1 & \text{if rand}(0,1) < 0.5 \\ 0 & \text{otherwise} \end{cases} & \text{if } rand\,(0,1) < MR \\ X_i & \text{otherwise} \end{cases} \quad (10)$$

Keterangan:

X = Solusi sebelum mutasi

Z = Solusi termutasi

X(i) dan Z(i) adalah elemen ke- i dalam solusi lama dan baru

i = 1, ...,D,

r = Nilai Random dengan range (minVal, maxVal)

MR = Mutation Rate

4. Fungsi fitness bertujuan untuk memaksimalkan tingkat akurasi klasifikasi (meminimalkan kesalahan klasifikasi) dan meminimalkan jumlah fitur terpilih. Berikut persamaannya:

$$fitness = \alpha\gamma_c + (1-\alpha)\frac{S}{N} \quad (11)$$

Keterangan:

α= Nilai antara 0-1

γ_c= Tingkat kesalahan pengklasifikasi

S = Jumlah fitur yang dipilih

N = Jumlah total fitur

Dalam penelitian ini (Khalid *et al.*, 2022) model pengklasifikasi yang digunakan yaitu KNN (dimana K=5), metode BCOVIDOA yang diusulkan diterapkan menggunakan 26 dataset yang berbeda dari UCI Machine Learning Repository, dari 26 dataset ini dipilih berdasarkan variasi ukuran dimensi (jumlah fitur). Dengan rincian ukuran dimensi kecil (9,11,13), sedang (64,91,256), dan dimensi besar (500,617,10000).

Algoritma BCOVIDOA ini dikomparasikan dengan beberapa algoritma seleksi fitur metaheusristik lainnya seperti GA, PSO, DE, WOA, WOASA, GWOPSO, HH, GWO, dan AOA (Khalid *et al.*, 2022). Setelah dilakukan pengujian dengan menggunakan 26 jenis dataset hasil akurasi rata-rata mencapai 92%, best fitness 0,0898, rata-rata fitness 0.0920, standar deviasi 0,0019, dan ukuran seleksi fitur rata-rata 147,15. Hasil yang diperoleh mengungkapkan efisiensi dari algoritma yang diusulkan serta menunjukkan kemampuan eksplorasi dan ekploitasi yang kuat.

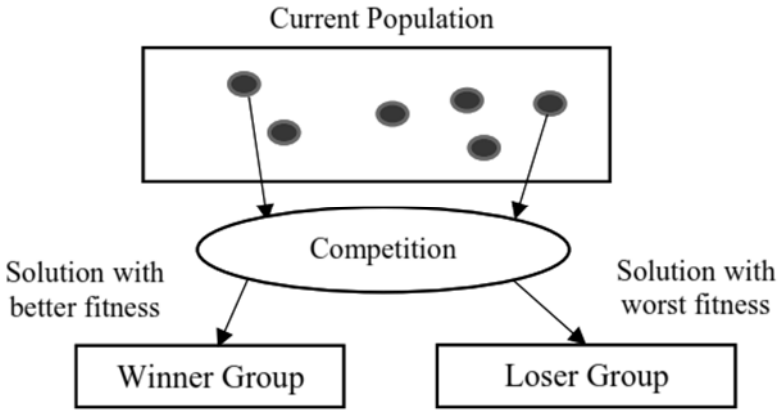
Hasil pengujian dari algoritma BCOVIDOA terhadap algoritma canggih lainnya menunjukkan bahwa algoritma BCOVIDOA mencapai akurasi maksimum pada 22 dari 26 dataset, namun GA mencapai akurasi maksimum hanya dalam 5 dari 26 dataset. Algoritma BCOVIDOA berada di posisi pertama dengan akurasi rata-rata 92,5 %, diikuti oleh algoritma GWOPSO dengan akurasi rata-rata 90%. BCOVIDOA sangat unggul dari algoritma lain dalam 19 dari 26 dataset.

Selain itu, hasil uji statistik rank-sum Wilcoxon mengungkapkan bahwa Algoritma BCOVIDOA berkinerja lebih baik daripada 9 sembilan algoritma canggih metaheuristik lainnya. Kurva konvergensi yang tervisualisasikan pada penelitian (Khalid *et al.*, 2022) membuktikan bahwa memiliki kecepatan konvergensi yang tinggi karena mencapai optimum global dengan cepat. Hanaa M. Hamza dkk (Khalid *et al.*, 2022) menyimpulkan bahwa BCOVIDOA lebih baik dari semua algoritma pembanding.

2.5 *Competition Strategy* dari Rival Genetic Algorithm (RGA)

RGA mengintegrasikan konsep kompetisi atau persaingan untuk membagi kromosom menjadi pemenang dan pecundang (kalah), serta memastikan induk berkualitas tinggi dalam proses seleksi (Too and Abdullah, 2021). Konsep RGA terinspirasi dari penelitian sebelumnya (Gu, Cheng and Jin, 2018) dan (Eshtay, Faris and Obeid, 2018) dan (Cheng and Jin, 2015) yang mengusulkan *Competitive Swarm Optimization* (CSO) yang telah menunjukkan bahwa penerapan *Competition Strategy* tidak hanya dapat meningkatkan kinerja sistem tetapi juga mengurangi kompleksitas komputasi.

Pada **Gambar 3** yang bersumber dari penelitian (Too and Abdullah, 2021) memberikan konsep *Competition Strategy* di RGA yaitu awalnya kromosom (solusi) dibagi secara acak menjadi pasangan *N/2*, simbol *N* adalah ukuran populasi, atau dengan penjelasan lain bahwa kromosom dipilih secara acak, dan berpasangan, dari populasi untuk berkompetisi. Selanjutnya dilakukan kompetisi antara dua kromosom pada masing-masing pasangan.



Gambar 5 Konsep *Competition Strategy* di RGA

RGA tidak secara langsung memindahkan pemenang ke generasi berikutnya, namun RGA membagi kromosom menjadi dua kelompok (pemenang dan kalah). Berdasarkan aturan kelangsungan hidup, kromosom yang memperoleh nilai fitness lebih baik diketahui sebagai pemenang, sehingga masuk kedalam kelompok pemenang. Sebaliknya kromosom yang memiliki nilai fitness rendah atau buruk masuk ke dalam kelompok kalah.

Dalam Algoritma Genetika konvensional, Parents dipilih secara random dari populasi menggunakan teknik seleksi *roulette wheel selection* atau *tournament selection.* Untuk seleksi *roulette wheel selection* probabilitas kromosom untuk menjadi Parents didasarkan pada nilai fitness yang diperoleh. Tetapi probabilitas dihitung untuk semua kromosom (N). Artinya, kromosom yang lemah (memiliki nilai fitness yang buruk) juga memiliki peluang untuk dipilih menjadi Parents. Selain itu juga terlalu banyak kromosom yang ikut serta dalam proses seleksi, sehingga dua kromosom yang lemah dapat menjadi Parents, oleh sebabnya akan mereproduksi kromosom (solusi) yang berkualitas rendah (Too and Abdullah, 2021).

Untuk mengatasi masalah tersebut, RGA memiliki skema operator seleksi baru untuk mengatasi masalah operasi seleksi pada operator seleksi GA konvensional, yaitu kromosom dibagi rata menjadi kelompok menang dan kelompok kalah berdasarkan hasil kompetisi nilai fitness, dengan ketentuan bahwa kelompok pemenang terdiri dari kromosom yang memiliki nilai fitness tinggi. Oleh karena itu, dengan memilih Parents dari kelompok pemenang, maka dapat dipastikan Parents yang terpilih berkualitas tinggi. Dalam RGA, Parents pertama dan kedua disimbolkan (X_{w1} dan X_{w2}) dipilih dari kelompok pemenang. Sedangkan Parent ketiga (X_l) dipilih dari kelompok yang kalah.

Setelah proses seleksi, berikutnya yaitu proses persilangan (crossover) dilakukan terhadap tiga Parents (X_{w1} , X_{w2} , X_l) untuk menghasilkan solusi baru (X_{new}). Persamaan utama Crossover sebagai berikut :

$$X_{new} = Crossover\ (X_{w1}\ ,\ X_{w2}\ ,\ X_l) \tag{12}$$

Strategi persilangan stochastic sederhana digunakan perdimensi untuk persilangan Parents (X_{w1} , X_{w2} , X_l) sebagai berikut :

$$X_{new}^{d} = \begin{cases} X_{w1}^{d}, \text{ if } r_1 \leq \frac{1}{3} \\ X_{w2}^{d}, \text{ if } \frac{1}{3} < r_1 \leq \frac{2}{3} \\ X_{l}^{d}, \text{ otherwise} \end{cases} \quad (13)$$

Keterangan:

X_{w1} , X_{w2} , X_l = Parents

r_1 = Vektor acak antara 0 dan 1

d = Dimensi ruang pencarian

Persamaan (13) crossover diatas dibuat dari kelompok menang dan kelompok kalah, Penelitian Too dan Abdullah (2021) meyakini bahwa dengan melakukan persilangan antara kelompok pemenang dan kelompok kalah, akan menghasilkan kromosom (solusi) berkualitas tinggi.

Pseudocode seleksi dan crossover pada RGA diilustrasikan dalam Algoritma 1 sebagai berikut.

Algoritma 1 : Selection and Crossover (RGA)

1) Randomly select two parents, X_{w1} and X_{w2} from winner group using roulette wheel selection
2) Randomly select a parents, X_1 from loser group using roulette wheel selection
3) Perform crossover between X_{w1} , X_{w2} , and X_1 as shown in (Persamaan 13)
4) Reproduce new solution & store it in the new population Z

Pseudocode RGA (Too and Abdullah, 2021) dalam Algoritma 2, tidak jauh berbeda dengan GA konvensional, populasi kromosom ***N*** diinisialisasi secara acak atau random, kemudian fitness setiap kromosom dihitung dan solusi terbaik ditentukan. Untuk setiap generasi, tingkat mutasi dihitung menggunakan persamaan (14) , setelah itu kromosom dipilih secara acak serta berpasangan dari populasi untuk kompetisi (perhatikan Gambar 5). Dari kompetisi tersebut, kromosom dengan nilai fitness yang lebih baik dipindahkan pada kelompok pemenang, sedangkan yang kalah ditempatkan pada kelompok kalah. Setelah itu dilakukan seleksi Parents (Induk) dan lakukan persilangan (crossover), seperti ilustrasi Algoritma 1. Kemudian, mutasi diterapkan pada setiap solusi yang baru

dihasilkan, dan nilai fitness dari solusi baru dievaluasi. Pada akhir setiap generasi, populasi baru digabungkan dengan populasi saat ini dan disortir menurut nilai fitnessnya. Kemudian, kromosom ***N*** terbaik disimpan untuk generasi berikutnya, dan kromosom terbaik global diperbarui. Algoritma diulang sampai jumlah maksimum generasi telah tercapai. Sehingga solusi optimal (subset fitur optimal) tercapai. Berikut ilustrasi Model Pseudocode Rival Genetic Algorithm (RGA).

Algoritma 2 : Rival Genetic Algorithm (RGA)

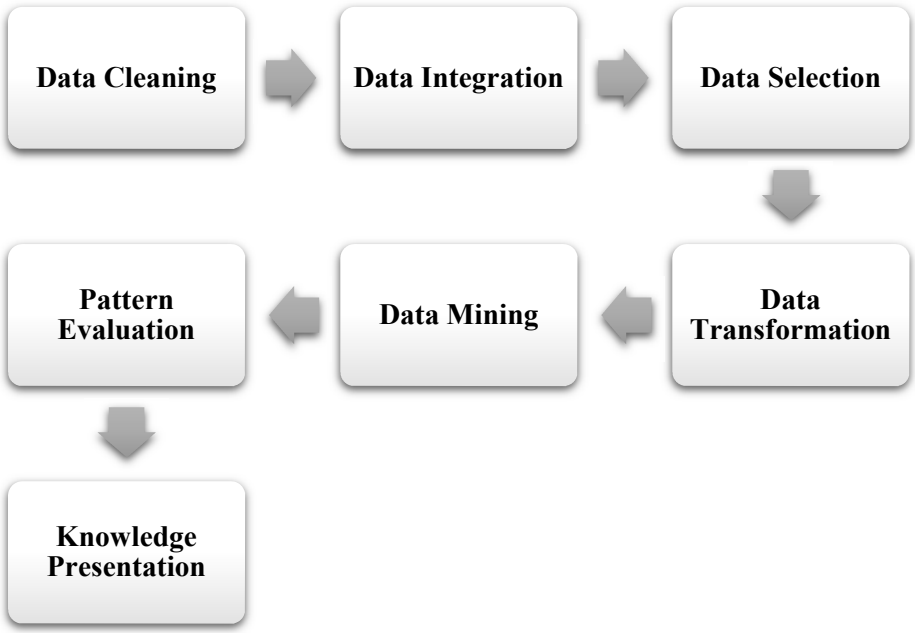
Input : Original feature set
Parameters: *N* and *MaxGene*
1) Initialize a population of chromosomes, *X*
2) Calculate the fitness values of chromosomes, *F(X)*
3) Set the global best chromosome as ***G***, set N_c =*N*
4) **for** g= 1 to *MaxGene*
5) Compute *MR* using
   // Competition Strategy //
6) **for** i= 1 to *N/2*
7) Random select two chromosomes, Xk and Xm
8) **if** F(Xk)better than F(Xm)
9) *Xw= Xk,Xl = Xm*
10) **else**
11) *Xw= Xm,Xl= Xk*
12) **end if**
13) Add Xw into winner group & add Xl into loser group
14) Remove Xk and Xm from the population
15) **next** i
   // Selection, Crossover & Mutation //
16) **for** i = 1 to N_c
17) **Perform selection and crossover as shown in (Algoritma 1)**
18) **for** d= 1 to D
19) **if** MR>=rand (0,1)
20) $Z^d_i = 1-Z^d_i$
21) **end if**
22) **next** d
23) **nex**t i
24) Calculate the fitness of new solutions
25) Merge the new population **Z** and current population **X**
26) Keep ***N*** best chromosomes for next generation
27) Update ***G*** if there is a better solution
28) **next** g
Output : Optimal feature subset

2.6 Data Mining

Menurut Qamar dan Raza (2020), data mining atau penggalian data merupakan inti dari seluruh proses analisis data. Hal ini dapat mencakup ekstraksi data dari sumber yang heterogen termasuk teks, video, angka dan gambar. Data

diekstraksi dari sumbernya, ditransformasikan dalam beberapa bentuk yang dapat dengan mudah diproses, dan akhirnya memuat data sehingga dapat melakukan pemrosesan. Namun, perlu diketahui bahwa seluruh proses tersebut memakan waktu dan membutuhkan banyak sumber daya, jadi, salah satu tujuan utamanya adalah untuk melakukan seluruh proses dengan efisien.

Definisi yang lain menurut Han *et.al* (2012) Secara analogi, data mining seharusnya lebih tepat dinamakan "penambangan pengetahuan dari data," yang penjelasannya terlalu panjang. Namun, dengan istilah yang lebih pendek, penambangan pengetahuan mungkin tidak mencerminkan penekanan pada penambangan dari data dalam jumlah besar. Namun demikian, penambangan adalah istilah yang jelas yang menggambarkan proses yang menemukan sekumpulan kecil bongkahan berharga dari sejumlah besar bahan mentah . Dengan demikian, istilah yang salah kaprah yang mengandung "data" dan "penambangan" menjadi populer . Selain itu, banyak istilah lain yang memiliki arti yang mirip dengan data mining - misalnya, penggalian pengetahuan dari data, ekstraksi pengetahuan, analisis data/pola, data arkeologi data, dan pengerukan data.



Gambar 6 Alur Proses Data mining

Proses Penemuan pengetahuan menurut Han *et.al* (2012) yang ditunjukkan pada **Gambar 4** sebagai urutan berulang dari langkah-langkah berikut:

- **Data Cleaning** (untuk menghilangkan noise dan data yang tidak konsisten).
- **Data Integration** (beberapa sumber data dapat digabungkan).
- **Data Selection** (data yang relevan dengan tugas analisis diambil dari database).
- **Data Transformation** (data ditransformasikan dan dikonsolidasikan ke dalam bentuk yang sesuai untuk penambangan dengan melakukan operasi ringkasan atau agregasi).
- **Data mining** (proses penting di mana metode cerdas diterapkan untuk mengekstrak pola data).
- **Pattern Evaluation** (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan).
- **Knowledge Presentation** (visualisasi dan teknik representasi pengetahuan digunakan untuk menyajikan pengetahuan yang telah ditambang kepada pengguna)

Langkah 1 sampai 4 adalah bentuk-bentuk yang berbeda dari prapemrosesan data, di mana data dipersiapkan untuk penambangan. Langkah penambangan data dapat berinteraksi dengan pengguna atau basis pengetahuan. Pola-pola yang pola yang menarik disajikan kepada pengguna dan dapat disimpan sebagai pengetahuan baru dalam basis pengetahuan.

Pandangan sebelumnya menunjukkan data mining sebagai salah satu langkah dalam proses penemuan pengetahuan, meskipun langkah yang penting karena mengungkap pola-pola yang tersembunyi untuk dievaluasi. Akan tetapi, dalam industri, media, dan lingkungan penelitian, istilah data mining sering digunakan untuk mengacu pada seluruh proses penemuan pengetahuan (mungkin karena istilahnya lebih pendek dari penemuan pengetahuan dari data). Oleh karena itu, kami mengadopsi pandangan yang luas tentang data mining fungsionalitas: Data mining adalah proses menemukan pola yang menarik dan dan pengetahuan dari sejumlah besar data. Sumber data dapat mencakup database, gudang data,

Web, repositori informasi lainnya, atau data yang dialirkan ke sistem secara dinamis.

2.7 Teknik Resampling

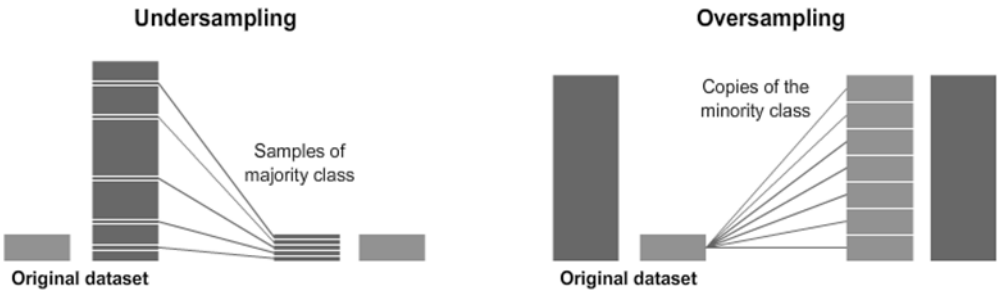
Teknik resampling adalah salah satu teknik preprocessing dimana distribusi data diseimbangkan kembali untuk mengurangi efek distribusi kelas tidak seimbang dalam proses pembelajaran (Beinecke and Heider, 2021). Teknik resampling secara luas digunakan untuk memecahkan masalah data yang tidak seimbang. Teknik ini dilakukan dengan mencoba menyeimbangkan data asli berdasarkan serangkaian algoritma sampling dengan menyesuaikan jumlah sampel dalam kelas yang berbeda, kemudian melatih data "seimbang" baru dengan mengadopsi algoritma klasifikasi (Azad *et al.*, 2022).

Ukuran kecil dari kelas minoritas dapat menyebabkan kemampuan algoritma pembelajaran *(learning algorithms*) untuk menemukan pola dalam distribusi kelas minoritas tidak dapat diandalkan. Namun dengan meningkatkan kelas minoritas dapat meningkatkan kemampuan algoritma pembelajaran menjadi lebih baik, karena bisa mengenali sampel kelas minoritas dari mayoritas (Muntasir Nishat *et al.*, 2022). Teknik Resampling adalah cara yang paling populer untuk mengatasi masalah ini.

Menurut Aftab and Matloob (2019) Pendekatan resampling dibagi menjadi tiga kategori :

- **RUS (Random Under-Sampling)**, menghapus beberapa kelas mayoritas dan membuat kelas menjadi sama, dataset yang seimbang dapat kehilangan informasi penting selama penghapusan instance dari kelas mayoritas sehingga bisa menjadi kinerja yang lebih buruk
- **ROS (Random Over Sampling)**, teknik ini menguramhi rasio ketidakseimbangan dalam dataset dengan menduplikasi instance di kelas minoritas. Pendekatan ini dapat menghasilkan kedua kelas yang seimbang namum masalah over-fitting dapat terjadi dan akan menurunkan kinerja pengklasidikasian.

- **SMOTE (Synthetic Minority Over-sampling Technique)**, teknik dilakukan dengan mensintesis sampel baru dari kelas minoritas untuk meyemimbangankan class dengan cara pembentukan instance yang



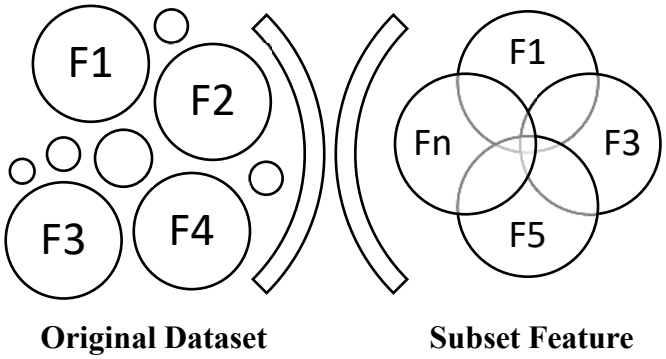
berdekatan

Gambar 7 Konsep Teknik Resampling

2.8 Seleksi Fitur

Menurut Qamar dan Raza (2020) seleksi fitur adalah Sebuah proses yang memilih fitur dari fitur yang diberikan subset tanpa mengubah atau kehilangan informasi. Jadi, kita dapat mempertahankan semantik data dalam proses transformasi.

Seleksi fitur, juga dikenal sebagai seleksi variabel, seleksi fitur, atau seleksi subset variabel, adalah teknik data mining yang menargetkan untuk memilih subset fitur yang optimal dari keseluruhan set fitur yang menghasilkan kinerja terbaik dalam hal definisi yang baik sesuai kriteria. Di sini, fitur mengacu pada fitur data, yang mewakili fungsi data ini dalam aspek tertentu. Karena seleksi fitur berkinerja baik dalam menyederhanakan model, mempersingkat waktu pelatihan, dan mengurangi varians model, peneliti dapat menafsirkan dan memahami pola model data dengan lebih mudah dengan menggunakan seleksi fitur (Rong, Gong and Gao, 2019).



Gambar 8 Konsep Seleksi Fitur

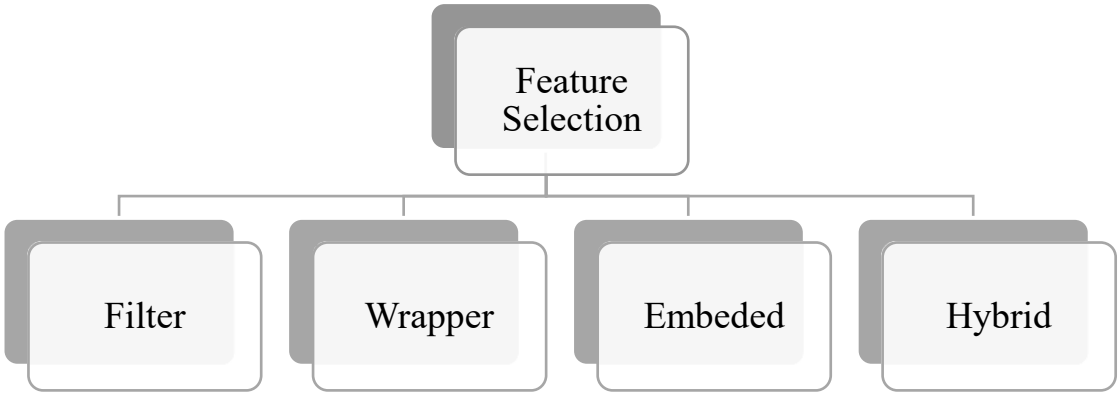
2.9 Metode Seleksi Fitur

Ada beberapa metode untuk melakukan seleksi fitur yang membantu memilih secara otomatis dari kumpulan data asli fitur yang paling efisien. *Wrapper*, *filter*, dan embedded merupakan tiga strategi seleksi fitur (Guo *et al.*, 2019). Literatur yang lain mengklasifikasikan proses seleksi fitur menjadi empat, yaitu metode *filter*, *wrapper*, embedded dan hybrid (Abiodun *et al.*, 2021).

Metode seleksi fitur adalah alat penemuan pengetahuan yang memberikan pemahaman tentang masalah melalui analisis fitur yang paling relevan. Seleksi fitur bertujuan untuk membangun pengklasifikasi yang lebih baik dengan membuat daftar fitur signifikan yang juga membantu mengurangi beban komputasi yang berlebihan (Khaire and Dhanalakshmi, 2022). Seleksi fitur memberikan banyak keuntungan: mengurangi ukuran data, mengurangi penyimpanan yang dibutuhkan, meningkatkan akurasi prediksi, menghindari overfitting, dan mengurangi waktu eksekusi dan pelatihan dari variabel yang mudah dipahami (Zebari *et al.*, 2020).

Metode *Filter* menerapkan beberapa analisis statistik pada satu set fitur untuk menyelesaikan permasalahan seleksi fitur tanpa menggunakan model pembelajaran (learning algorithm).Diantaranya chi-square test, information gain

dan correlation coefficient (Moradi and Gholampour, 2016). Oleh karena itu, metode ini biasanya tidak membutuhkan waktu komputasi yang banyak.



Gambar 9 Jenis Metode Seleksi Fitur

Metode *Wrapper* menggunakan algoritma pembelajaran untuk melakukan evaluasi terhadap subset fitur melalui proses pencarian. Dengan kata lain, metode *wrapper* adalah proses pencarian berulang dimana hasil dari algortima pembelajaran pada setiap iterasi digunakan untuk memandu proses pencarian (Moradi and Gholampour, 2016).

Metode Embeded menggunakan algoritma dengan metode seleksi fitur bawaannya sendiri. Ini mirip dengan metode *wrapper* dimana pengklasifikasi yang sama digunakan dalam memilih fitur pada fase evaluasi. Namun, menggunakan pengklasifikasi dalam metode embeded dicapai dengan biaya komputasi yang lebih rendah daripada metode *wrapper* (Abiodun *et al.*, 2021).

Sedangkan Metode Hybrid adalah kombinasi dari metode *filter* dan *wrapper* (Moradi and Gholampour, 2016). Metode hybrid berfokus kepada penggabungan *filter* dan *wrapper* untuk mencapai performa terbaik yaitu dengan menggunakan algoritma pembelajaran tertentu dengan waktu kompleksitas mirip dengan metode *filter* (Moradi and Gholampour, 2016). Teknik hybrid menggunakan lebih dari satu strategi untuk memilih fitur untuk membuat himpunan bagian. Ini menggabungkan beberapa pendekatan untuk mendapatkan

subset fitur terbaik daripada menggunakan metode independen. Dalam pendekatan hibrida, dua metode dapat digabungkan secara logis, misalnya metode pembungkus dan *filter*. Ini dimulai dengan metode *filter* yang digunakan untuk membuat subset fitur, diikuti dengan metode *wrapper* yang digunakan untuk memilih fitur dari subset (Wah *et al.*, 2018).

Metode hybrid dibangun di atas intuisi menciptakan model yang efektif dan efisien dengan menggabungkan metode yang lebih lemah, sehingga, dikenal istilah hybrid. Metode hybrid dapat melakukan seleksi fitur dan pelatihan model secara bersamaan. Akurasi dan kinerja yang tinggi, kompleksitas komputasi yang optimal, model yang kuat dan fleksibel adalah beberapa manfaat yang dinikmati dari metode hybrid (Abiodun *et al.*, 2021).

2.10 Klasifikasi

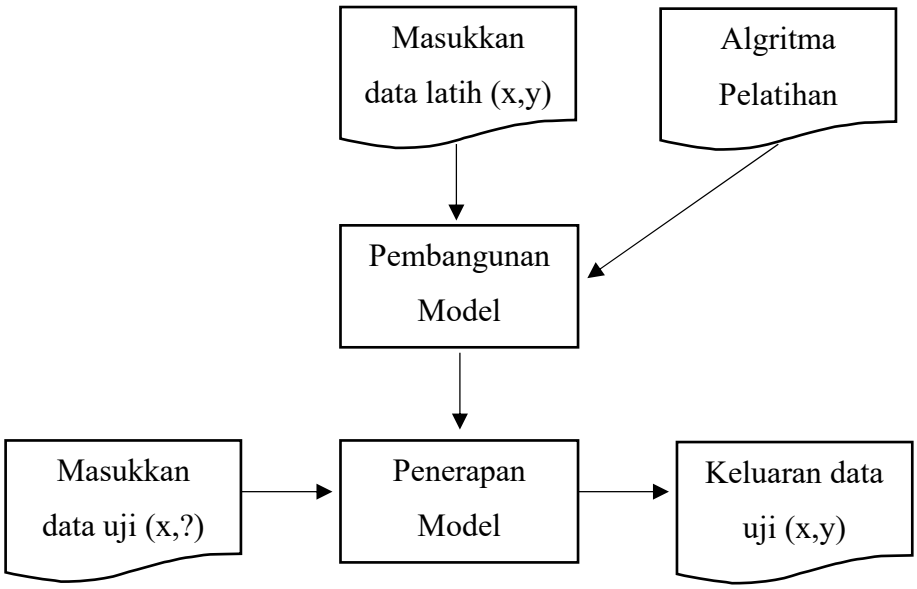
Teknik klasifikasi dalam penggalian data yang mampu memproses data dalam jumlah yang besar. Hal ini dapat memprediksi kategoris label kelas dan data mengklasifikasikan berdasarkan set pelatihan dan kelas label dan karenanya dapat digunakan untuk mengelompokkan data yang baru tersedia. Dengan demikian dapat dijelaskan sebagai bagian tak terelakkan dari penggalian data dan mendapatkan popularitas yang lebih (Durugkar *et al.*, 2022).

Klasifikasi adalah teknik yang digunakan secara luas di berbagai bidang, termasuk data pertambangan, yang tujuannya adalah untuk mengklasifikasikan set besar objek ke dalam kelas yang telah ditetapkan, dijelaskan oleh satu set fitur, menggunakan metode pembelajaran yang terawasi. Karena ledakan pertumbuhan antara basis data bisnis dan ilmiah, penggalian aturan efisiensi klasifikasi dari basis data tersebut sangat penting (Khurma *et al.*, 2022).

Lan *et al.*, (2018) mengungkapkan bahwa analisis klasifikasi adalah menganalisis data dalam basis data, untuk membuat deskripsi yang akurat atau membangun model yang akurat atau menambang aturan pengklasifikasian untuk setiap kategori, dan kemudian menggunakan aturan untuk mengklasifikasikan catatan dalam basis data lainnya.

Klasifikasi dapat didefinisikan secara detail sebagai suatu pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target f yang menetapkan

setiap vektor (set fitur) x ke dalam satu dari sejumlah label kelas y yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori (Prasetyo, 2014). Model dalam klasifikasi mempunyai arti yang sama dengan blackbox, di mana ada suatu model yang menerima masukan kemudian mampu melakukan pemikiran terhadap masukan tersebut dan memberikan jawaban sebagai keluaran dari hasil pemikirannya. Kerangka kerja klasifikasi ditunjukkan pada **Gambar 8**. Dalam gambar tersebut, disediakan sejumlah data latih (x, y) untuk digunakan sebagai data membangun model, kemudian menggunakan model tersebut untuk memprediksi kelas dari data uji (x,?) sehingga data uji (x, ?) diketahui kelas y yang seharusnya.



Gambar 10 Kerangka Kerja Klasifikasi

Model yang sudah dibangun pada saat pelatihan kemudian dapat digunakan untuk memprediksi label kelas dari data baru yang belum diketahui label kelasnya. Dalam pembangunan model selama proses pelatihan tersebut diperlukan adanya suatu algoritma untuk membangunnya yang disebut sebagai algoritma pelatihan (learning algorithm). Kerangka kerja seperti yang ditunjukkan pada **Gambar 10** meliputi dua langkah proses yaitu induksi dan deduksi. Induksi merupakan suatu langkah untuk membangn klasifikasi dari data latih yang diberikan atau disebut juga dengan proses pelatihan, sedangkan deduksi

merupakan suatu langkah untuk menerapkan model tersebut pada data uji sehingga data uji dapat diketahui kelas yang sesungguhnya atau disebut juga dengan proses prediksi.

Berdasarkan cara pelatihan, algoritma-algoritma klasifikasi dapat dibagi menjadi dua macam, yaitu lazy learner dan eager learner. Algoritma-algoritma yang masuk kategori lazy learner hanya sedikit melakukan pelatihan (atau bahkan tidak sama sekali). Algoritma-algoritma ini hanya menyimpan sebagian atau seluruh data latih, kemudian menggunakan data latih tersebut ketika proses prediksi. Hal ini mengakibatkan proses prediksi menjadi lama karena model harus membaca kembali semua data latihnya untuk dapat memberikan keluaran label kelas dengan benar pada data uji yang diberikan. Kelebihan dari algoritma seperti ini adalah proses pelatihan berjalan dengan cepat. Algoritma-algoritma klasifikasi yang masuk kategori ini di antaranya adalah rote classifier, K-Nearest Neighbor (K-NN), Fuzzy K-Nearest Neighbor (FK-NN), regresi linear dan sebagainya.

Sedangkan algoritma-algoritma yang masuk kategori eager learner didesain untuk melakukan pembacaan/pelatihan/pembelajaran pada data latih untuk dapat memetakan dengan benar setiap vektor masukan ke label kelas keluarannya sehingga di akhir proses pelatihan, model sudah dapat melakukan pemetaan dengan benar semua data latih ke label kelas keluarannya. Setelah proses pelatihan tersebut selesai, maka model (biasanya berupa bobot atau sejumlah nilai kuantitatif tertentu) disimpan sebagai memori, sedangkan semua data latihnya dibuang. Proses prediksi dilakukan menggunakan model yang tersimpan dan tidak melibatkan data latih sama sekali. Cara ini mengakibatkan proses prediksi dapat berjalan dengan cepat, namun harus dibayar dengan proses pelatihan yang lama. Algoritma-algoritma klasifikasi yang masuk kategori ini di antaranya Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree, Bayesian dan sebagainya.

2.11 Kompleksitas Algoritma

Sebuah algoritma tidak saja harus benar tetapi juga harus efisien. Keefisienan algoritma diukur dari waktu eksekusi algoritma dan kebutuhan ruang memori. Algoritma yang efisien adalah algoritma yang meminimumkan kebutuhan waktu dan ruang. Kebutuhan waktu dan ruang suatu algoritma bergantung pada ukuran masukan (n), yang menyatakan jumlah data yang diproses. Keefisienan algoritma dapat digunakan untuk menilai algoritma yang bagus dari sejumlah algoritma penyelesaian masalah (Mauluddin, Ikbal and Nursikuwagus, 2020).

Kompleksitas algoritma terdiri dari dua macam yaitu kompleksitas waktu dan kompleksitas ruang. Kompleksitas waktu, dinyatakan oleh *T(n)*, diukur dari jumlah komputasi yang dibutuhkan untuk menjalankan algoritma sebagai fungsi dari ukuran masukan *n*, di mana ukuran masukan *(n)* merupakan jumlah data yang diproses oleh sebuah algoritma. Sedangkan kompleksitas ruang, *S(n)*, diukur dari memori yang digunakan oleh struktur data yang terdapat di dalam algoritma sebagai fungsi dari masukan *n*. Dengan memenfaatkan kompleksitas waktu atau kompleksitas ruang, laju peningkatan waktu dan algoritma dapat ditentukan seiring dengan meningkatnya ukuran masukan *n* (Salecha, 2021).