OPTIMASI KINERJA ALGORITMA *MACHINE LEARNING* PADA KLASIFIKASI PENYAKIT CACAR MONYET MENGGUNAKAN *INFORMATION GAIN* DAN GRIDSEARCHCV

Steven Joses, Donata Yulvida, Genrawan Hoendarto
Informatika, Universitas Widya Dharma Pontianak
Jalan HOS Cokroaminoto No.445 Pontianak, Indonesia
stevenjoses@widyadharma.ac.id

ABSTRAK

Penyakit cacar monyet merupakan infeksi virus zoonotik dari genus *Orthopoxvirus* yang memiliki kemiripan gejala dengan beberapa penyakit lain sehingga menyulitkan proses identifikasi dini. Permasalahan yang dihadapi adalah rendahnya ketepatan diagnosis apabila hanya mengandalkan pengamatan klinis tanpa dukungan metode berbasis data. Penelitian ini bertujuan untuk membangun model klasifikasi penyakit cacar monyet serta menganalisis pengaruh seleksi fitur dan optimasi parameter terhadap performa model. Metode yang digunakan meliputi *Support Vector Machine*, *Random Forest*, dan XGBoost dengan penerapan seleksi fitur *Information Gain* serta optimasi parameter menggunakan GridSearchCV. Tahapan penelitian meliputi pengujian model dengan parameter *default*, penambahan seleksi fitur, dan optimasi parameter. Evaluasi dilakukan menggunakan metrik akurasi dan *F1-Score*. Hasil penelitian menunjukkan bahwa penerapan seleksi fitur dan optimasi parameter mampu meningkatkan performa klasifikasi. Model terbaik diperoleh menggunakan XGBoost dengan akurasi sebesar 71,3% dan F1-Score sebesar 81,4%.

Kata kunci : *Penyakit Cacar Monyet, Random Forest, SVM, XGBoost, Information Gain, GridSearchCV*

1. PENDAHULUAN

Penyakit cacar monyet merupakan infeksi virus zoonotik yang disebabkan oleh virus cacar monyet dan ditandai dengan munculnya ruam yang menyerupai penyakit cacar [1]. Penyakit ini umumnya diawali dengan demam, diikuti oleh munculnya lesi kulit yang khas dengan distribusi tertentu pada tubuh. Secara virologis, virus cacar monyet termasuk dalam keluarga *Poxviridae*, subfamili *Chordopoxvirinae*, dan genus *Orthopoxvirus*. Genus ini juga mencakup beberapa virus lain seperti *variola*, *vaccinia*, *cowpox*, dan *camelpox*. Virus ini merupakan virus DNA untai ganda yang memiliki kemiripan genetik dan antigenik yang tinggi dengan virus-virus dalam genus yang sama, sehingga memungkinkan terjadinya kekebalan silang. Vaksinasi terhadap cacar diketahui dapat memberikan perlindungan parsial terhadap infeksi cacar monyet [2].

Gejala cacar monyet pada manusia memiliki kemiripan dengan gejala cacar air, meskipun tingkat keparahannya umumnya lebih ringan. Gejala awal yang sering muncul meliputi demam, sakit kepala, nyeri otot, dan kelelahan [3]. Kemiripan gejala ini seringkali menyulitkan proses identifikasi penyakit secara dini apabila hanya mengandalkan pengamatan klinis secara konvensional. Pemanfaatan teknologi kecerdasan buatan melalui analisis data gejala yang dialami pasien dapat menjadi salah satu pendekatan yang efektif dalam membantu proses identifikasi penyakit cacar monyet. Seiring dengan meningkatnya jumlah kasus serta kompleksitas faktor epidemiologis yang berkontribusi terhadap penyebaran cacar monyet, diperlukan penerapan pendekatan analisis berbasis data yang lebih komprehensif guna mendukung upaya deteksi dini, pemantauan penyakit, serta pengambilan keputusan dalam bidang kesehatan masyarakat.

Salah satu bentuk kecerdasan buatan yang dapat dimanfaatkan untuk mengidentifikasi penyakit cacar monyet adalah *machine learning*. *Machine learning* merupakan metode yang memungkinkan sistem komputer mempelajari pola dari data untuk menghasilkan prediksi atau keputusan secara otomatis. Dalam bidang kesehatan, teknologi ini memiliki berbagai manfaat yang signifikan. Beberapa di antaranya meliputi identifikasi penyakit, diagnosis penyakit, serta prediksi perkembangan penyakit. Penerapan *machine learning* juga dapat dimanfaatkan dalam pengelolaan rekam medis secara cerdas, analisis citra medis, serta berbagai penerapan lainnya [3].

Terdapat beberapa penelitian yang membahas mengenai diagnosa penyakit cacar monyet. Pada penelitian [4] dilakukan penelitian untuk membangun model klasifikasi *machine learning* dalam diagnosis penyakit cacar monyet. Metode yang digunakan adalah XGBoost dan *Random Forest* (RF) untuk mengklasifikasikan pasien terinfeksi dan tidak terinfeksi berdasarkan gejala klinis. Hasil penelitian menunjukkan bahwa XGBoost memberikan performa terbaik dengan akurasi 68%, presisi 69%, recall 89%, dan F1-score 78%. Sementara itu, RF menghasilkan performa yang sedikit lebih rendah dengan nilai AUC sebesar 0,60 dalam evaluasi ROC.

Pada penelitian lain [5] menerapkan metode *Support Vector Machine* (SVM) untuk klasifikasi penyakit cacar monyet. Model dibangun menggunakan kernel *Radial Basis Function* (RBF) dengan eksplorasi parameter C (*cost*) dan γ (*gamma*). Hasil penelitian menunjukkan parameter terbaik pada C = 10 dan γ = 1. Evaluasi menggunakan *confusion*

matrix dengan rasio data 90:10 menghasilkan akurasi sebesar 65%. Hasil tersebut menunjukkan bahwa metode SVM dapat digunakan untuk klasifikasi cacar monyet, namun peningkatan jumlah dan keseimbangan data masih diperlukan untuk meningkatkan performa model.

Pada penelitian lainnya [3] dilakukan klasifikasi pasien yang terinfeksi virus cacar monyet menggunakan metode *machine learning*. Metode yang diterapkan meliputi beberapa algoritma, yaitu *K-Nearest Neighbor* (KNN), SVM, RF, dan *Gradient Boosting* (GB). Hasil penelitian menunjukkan bahwa GB memperoleh performa terbaik dengan akurasi 71%, diikuti oleh SVM 69%, RF 68%, dan KNN 63%. Hasil tersebut menunjukkan bahwa metode *machine learning* dapat digunakan untuk membantu proses klasifikasi penyakit cacar monyet.

Berdasarkan penelitian yang telah dilakukan sebelumnya, penelitian ini mengusulkan penerapan teknik seleksi fitur *Information Gain* (IG) untuk meningkatkan kinerja model *machine learning* dalam klasifikasi penyakit cacar monyet. Seleksi fitur berdasarkan nilai *Information Gain* digunakan untuk menentukan fitur-fitur dengan peringkat tertinggi yang memiliki pengaruh besar terhadap proses klasifikasi. Pengembangan model dilakukan dengan memanfaatkan fitur-fitur hasil seleksi tersebut sehingga hanya atribut yang paling relevan yang digunakan dalam proses pembelajaran model.

2. TINJAUAN PUSTAKA

2.1. Penyakit Cacar Monyet

Setelah terjadinya pandemi COVID-19 pada tahun 2020 yang memberikan dampak besar terhadap kesehatan masyarakat global, dunia kembali dihadapkan pada ancaman penyakit menular lainnya, yaitu cacar monyet. Penyakit ini menjadi perhatian karena berpotensi menimbulkan wabah serta berdampak pada sistem kesehatan masyarakat di berbagai negara. Cacar monyet merupakan penyakit zoonosis, yaitu penyakit yang dapat ditularkan dari hewan ke manusia melalui kontak langsung dengan hewan terinfeksi, cairan tubuh, maupun lesi kulit. Penyakit ini disebabkan oleh *Monkeypox virus* yang termasuk dalam genus *Orthopoxvirus*. Virus tersebut memiliki hubungan kekerabatan dengan *Variola virus*, yaitu virus penyebab penyakit cacar pada manusia [5].

Kasus pertama cacar monyet pada manusia dilaporkan pada tahun 1970 di Afrika, dalam konteks program pengawasan dan pemberantasan cacar yang sedang berlangsung pada masa itu. Kasus tersebut terjadi pada seorang anak laki-laki berusia sembilan bulan yang mengalami demam dan dua hari kemudian muncul ruam sentrifugal, yaitu ruam dengan distribusi lesi yang dominan pada lengan dan kaki. Antara September 1970 hingga Maret 1971, enam kasus tambahan berhasil diidentifikasi di beberapa negara Afrika Barat. Sebagian besar pasien adalah anak-anak kecil dan tidak pernah menerima vaksinasi cacar [2].

Sejak penghentian program vaksinasi cacar global pada tahun 1980, kekebalan populasi terhadap virus dalam genus *Orthopoxvirus* mengalami penurunan secara bertahap. Kondisi ini berkontribusi terhadap peningkatan kasus cacar monyet di beberapa wilayah endemik, termasuk peningkatan bertahap kasus di Republik Demokratik Kongo serta kemunculan kembali wabah di Nigeria pada tahun 2017. Selain menurunnya kekebalan kelompok, faktor lain yang berperan antara lain meningkatnya aktivitas perburuan dan konsumsi satwa liar, serta ekspansi urbanisasi yang mendorong perambahan hutan dan rawa sebagai habitat alami reservoir virus sehingga meningkatkan kemungkinan interaksi antara manusia dan hewan pembawa virus. Meskipun demikian, perhatian global terhadap wabah cacar monyet di Afrika relatif terbatas, sehingga penelitian dan respons internasional terhadap penyakit ini masih belum optimal [1].

2.2. Information Gain

Information Gain (IG) merupakan salah satu metode seleksi fitur yang banyak digunakan dalam proses pemilihan atribut pada data [6]. Metode ini memanfaatkan konsep entropi untuk menentukan atribut atau *term* yang paling optimal. IG mengukur seberapa besar informasi yang diberikan oleh suatu fitur terhadap proses klasifikasi. Semakin tinggi nilai IG yang dimiliki oleh suatu fitur, maka fitur tersebut dianggap semakin relevan dalam membedakan kelas.

Fitur yang memiliki nilai IG lebih besar dinilai memiliki tingkat signifikansi dan kontribusi yang lebih tinggi dalam proses pemodelan [7]. Dengan demikian, metode ini dapat membantu dalam mengidentifikasi serta memilih atribut yang paling relevan untuk digunakan dalam proses pembangunan model. Pemilihan fitur yang tepat juga dapat meningkatkan kinerja model serta mengurangi kompleksitas data yang digunakan.

2.3. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) merupakan salah satu algoritma *supervised learning* yang banyak digunakan dalam *machine learning*. Algoritma ini menerapkan pendekatan *ensemble learning* yang dikembangkan berdasarkan *metode Gradient Boosting* [8].

XGBoost merupakan algoritma yang membangun model secara bertahap dengan tujuan memperbaiki kesalahan yang dihasilkan oleh model sebelumnya. Proses pembelajaran ini memanfaatkan teknik optimasi berbasis *gradien* dan *hessian* untuk meningkatkan kinerja model. Melalui pendekatan tersebut, setiap iterasi model baru berusaha meminimalkan kesalahan prediksi dari model sebelumnya. Metode ini juga dirancang agar mampu bekerja secara efisien dalam proses komputasi [4].

2.4. Random Forest

Random Forest (RF) merupakan algoritma *machine learning* yang bekerja dengan menggabungkan sejumlah *decision tree* yang dibangun secara acak dari data pelatihan. Setiap pohon menghasilkan prediksi yang kemudian digabungkan untuk memperoleh hasil prediksi akhir yang lebih akurat. Salah satu keunggulan utama RF adalah kemampuannya dalam menangani data yang berukuran besar dan kompleks tanpa mudah mengalami *overfitting*. Selain itu, algoritma ini juga mampu memberikan hasil prediksi yang lebih stabil [4].

Pada RF, nilai keluaran untuk setiap data diperoleh dari rata-rata prediksi yang dihasilkan oleh beberapa *decision tree*. Proses ini diawali dengan pembentukan *bootstrap sample* dari *dataset* asli untuk menghasilkan sejumlah sampel data pelatihan. Dari setiap sampel tersebut kemudian dikembangkan pohon regresi tanpa proses pemangkasan (*unpruned tree*). Nilai prediksi akhir diperoleh dengan menghitung rata-rata hasil prediksi dari seluruh pohon yang terbentuk. Selain itu, tingkat kesalahan model dapat diperkirakan menggunakan metode *out-of-bag prediction* yang dihitung dari data pelatihan yang tidak termasuk dalam sampel bootstrap [9].

2.5. Support Vector Machine

Support Vector Machine (SVM) merupakan teknik pembelajaran mesin yang digunakan untuk pengenalan pola [5]. Metode ini bekerja dengan cara menganalisis data dan membangun model yang mampu memisahkan data ke dalam beberapa kelas. SVM dikenal memiliki kemampuan yang baik dalam menangani data berdimensi tinggi serta menghasilkan model yang memiliki tingkat generalisasi yang baik.

Konsep klasifikasi dengan SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik. *Hyperplane* tersebut berfungsi sebagai pemisah optimal antara dua kelas data yang ditentukan dengan memaksimalkan nilai margin dari *hyperplane* [10].

2.6. GridSearchCV

Salah satu metode yang dapat digunakan dalam proses pengambilan keputusan untuk menentukan parameter terbaik adalah GridSearchCV. GridSearchCV merupakan teknik yang sering digunakan dalam proses optimasi *hyperparameter* karena penerapannya relatif sederhana dan sistematis. Metode ini bekerja dengan cara menguji seluruh kombinasi parameter yang telah ditentukan sebelumnya untuk menemukan konfigurasi yang paling optimal. Dengan pendekatan tersebut, setiap kemungkinan parameter akan dievaluasi terhadap kinerja model [10].

2.7. Evaluasi Model

Untuk mengevaluasi model *machine learning* yang telah diterapkan, umumnya digunakan beberapa metrik pengukuran kinerja seperti akurasi, presisi, dan *recall*. Ketiga metrik tersebut digunakan untuk menilai seberapa baik model dalam melakukan proses klasifikasi. Akurasi menunjukkan tingkat kedekatan antara hasil prediksi yang dihasilkan oleh sistem dengan nilai sebenarnya pada data aktual [11].

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

Keterangan :
TP = Jumlah *True Positive*
TN = Jumlah *True Negative*
FP = Jumlah *False Positive*
FN = Jumlah *False Negative*

Precision merupakan metrik evaluasi yang digunakan untuk mengukur tingkat ketepatan model dalam memprediksi data positif [12].

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

Recall merupakan metrik evaluasi yang digunakan untuk menilai sejauh mana model mampu mengidentifikasi seluruh data positif yang relevan [12].

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

F1-score merupakan metrik evaluasi yang digunakan untuk merepresentasikan rata-rata tertimbang antara nilai *precision* dan *recall* [12].

$$F1\ Score = \frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

2.8. Penelitian Terdahulu

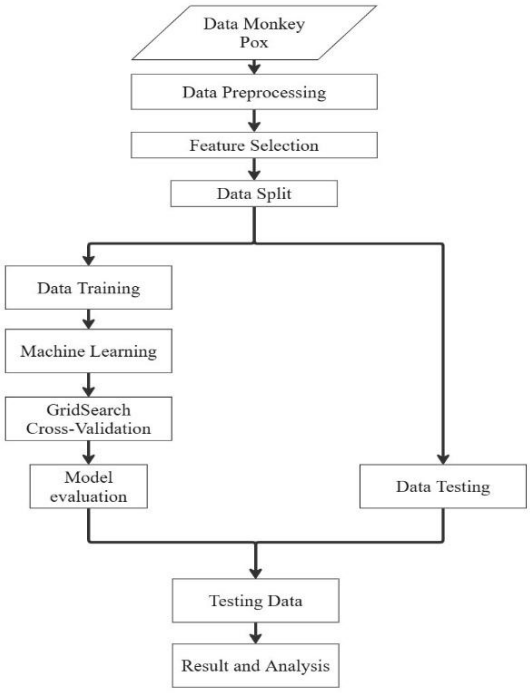
Penelitian sebelumnya [4] yang dilakukan oleh Mohammad Dito Dwi Krisna dan Firman Noor Hasan mengembangkan model klasifikasi penyakit cacar monyet menggunakan algoritma XGBoost dan RF berdasarkan data gejala klinis pasien. *Dataset* yang digunakan berasal dari *Kaggle* dengan jumlah 25.000 data pasien yang memuat beberapa atribut gejala seperti *systemic illness*, *rectal pain*, *sore throat*, *penile edema*, *oral lesions*, *swollen tonsils*, dan infeksi menular seksual lainnya. Proses penelitian meliputi tahapan *exploratory data analysis* (EDA), *preprocessing data*, pembagian data latih dan data uji dengan rasio 80:20, serta pembangunan model klasifikasi menggunakan kedua algoritma tersebut. Evaluasi model dilakukan menggunakan beberapa metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa XGBoost memiliki performa yang sedikit lebih baik dibandingkan RF dengan akurasi sebesar 68%, *precision* 69%, *recall* 89%, dan *F1-score* 78%, sedangkan kedua model memiliki nilai AUC sebesar 0,60 yang menunjukkan kemampuan klasifikasi pada tingkat moderat.

Penelitian sebelumnya [5] yang dilakukan oleh Wendy Anugrah et al menggunakan metode SVM untuk melakukan klasifikasi penyakit cacar monyet dengan memanfaatkan data gejala pasien. *Dataset* yang digunakan berasal dari *Kaggle* dengan jumlah 5000 data yang memuat beberapa gejala seperti *systemic illness*, *rectal pain*, *sore throat*, *oral lesions*, *penile oedema*, *solitary lesions*, *swollen tonsils*, dan *sexually transmitted infection*. Pada tahap *preprocessing*, atribut HIV *infection* dihapus karena tidak memiliki hubungan ilmiah yang kuat dengan penularan cacar monyet, sehingga diperoleh delapan variabel gejala yang digunakan dalam proses pemodelan. Proses klasifikasi dilakukan menggunakan SVM dengan kernel *Radial Basis Function* (RBF) untuk menangani pemisahan data yang bersifat nonlinier, dengan optimasi parameter menggunakan GridSearchCV pada nilai C dan *gamma*. Hasil pengujian menunjukkan bahwa model terbaik diperoleh pada parameter C = 10 dan γ = 1 dengan rasio data latih dan uji 90:10, yang menghasilkan akurasi sebesar 65%, dengan nilai *precision* 67%, *recall* 59%, dan *F1-score* 63%. Hasil tersebut menunjukkan bahwa metode SVM dapat digunakan untuk klasifikasi penyakit cacar monyet, namun peningkatan jumlah serta keseimbangan data masih diperlukan untuk meningkatkan performa model.

Penelitian sebelumnya [3] oleh Febri Aldi et al. melakukan klasifikasi penyakit cacar monyet menggunakan beberapa algoritma *machine learning* dengan *dataset* dari *Kaggle* yang berjumlah 25.000 data pasien. Analisis data dilakukan dengan menggunakan koefisien korelasi untuk melihat hubungan antar fitur serta jumlah variabel target pada *dataset*. Proses klasifikasi dilakukan menggunakan beberapa algoritma, yaitu *K-Nearest Neighbor* (KNN), SVM, RF, dan *Gradient Boosting* (GB). Hasil penelitian menunjukkan bahwa algoritma *Gradient Boosting* menghasilkan performa terbaik dengan tingkat akurasi sebesar 71%, diikuti oleh SVM sebesar 69%, RF sebesar 68%, dan KNN sebesar 63%. Hasil tersebut menunjukkan bahwa metode *machine learning* dapat digunakan untuk membantu proses analisis dan identifikasi penyakit cacar monyet sehingga dapat mendukung pengambilan keputusan di bidang kesehatan.

3. METODE PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan proses yang terstruktur untuk menghasilkan model klasifikasi yang optimal. Tahapan penelitian meliputi pengumpulan *dataset*, proses *preprocessing*, seleksi fitur menggunakan metode IG, pembagian data menjadi data latih dan data uji, pembangunan model *machine learning*, serta optimasi parameter menggunakan GridSearchCV. Secara keseluruhan tahapan penelitian yang dilakukan dapat dilihat pada alur penelitian pada Gambar 3.1.



Gambar 1. Alur penelitian

3.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Monkey-Pox Patients Dataset* yang diperoleh dari *Kaggle*. *Dataset* ini merupakan data sintetis yang dibuat berdasarkan penelitian yang dipublikasikan oleh The BMJ mengenai karakteristik klinis dan presentasi baru penyakit cacar monyet pada manusia selama wabah tahun 2022. *Dataset* tersebut berisi 25.000 data pasien yang disimpan dalam format CSV, dimana setiap data pasien memiliki sejumlah fitur yang merepresentasikan kondisi atau gejala yang dialami.

Beberapa fitur yang terdapat dalam dataset antara lain *Systemic Illness*, *Rectal Pain*, *Sore Throat*, *Penile Oedema*, *Oral Lesions*, *Solitary Lesion*, *Swollen Tonsils*, *HIV Infection*, dan *Sexually Transmitted Infection*. Selain itu, *dataset* juga memiliki variabel target yaitu *MonkeyPox* yang menunjukkan apakah pasien terindikasi menderita cacar monyet atau tidak. Data pada *dataset* ini sebagian besar berbentuk *boolean* dan kategorikal, sehingga sebelum digunakan dalam proses pemodelan *machine learning* perlu dilakukan tahap pra-pemrosesan data.

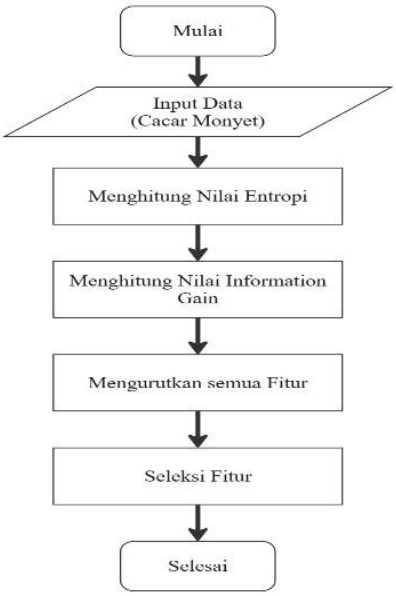
3.2. Data Preprocessing

Tahap data *preprocessing* merupakan tahap awal sebelum melakukan analisis data. Pada penelitian ini, proses *preprocessing* dilakukan untuk memastikan bahwa data yang digunakan memiliki kualitas yang baik sebelum digunakan dalam proses pemodelan *machine learning* untuk klasifikasi penyakit cacar monyet. Langkah pertama yang dilakukan adalah melakukan pemeriksaan terhadap *missing value* pada *dataset*. Pemeriksaan ini bertujuan untuk mengetahui

apakah terdapat data yang kosong atau tidak lengkap pada setiap atribut sehingga dapat memastikan konsistensi data yang digunakan dalam proses analisis.

Setelah dilakukan pemeriksaan *missing value*, tahap selanjutnya adalah melakukan transformasi data kategorikal menjadi data numerik. Proses ini dilakukan karena algoritma *machine learning* umumnya memerlukan data dalam bentuk numerik. Oleh karena itu, nilai-nilai kategorikal pada setiap fitur diubah menjadi bentuk angka melalui proses pengkodean (*encoding*). Proses ini bertujuan untuk mengelompokkan setiap kategori ke dalam representasi numerik sehingga data dapat digunakan secara optimal dalam proses pelatihan model klasifikasi.

3.3. Seleksi Fitur



Gambar 2. Metode Information Gain

Tahap seleksi fitur bertujuan untuk mengidentifikasi subset fitur yang paling relevan dan memberikan kontribusi signifikan dalam memprediksi variabel target. Pada penelitian ini, pendekatan yang digunakan adalah metode pengukuran informasi yaitu *Information Gain* (IG) untuk mengevaluasi tingkat kepentingan setiap fitur terhadap variabel target. IG digunakan untuk mengukur seberapa besar informasi yang dapat diberikan oleh suatu fitur dalam membantu proses klasifikasi terhadap kelas target.

Penggunaan nilai tertinggi dari IG memungkinkan penelitian ini untuk mengidentifikasi fitur-fitur yang paling informatif dan memiliki pengaruh signifikan terhadap variabel target. Dengan melakukan seleksi fitur menggunakan metode ini, dimensi data dapat dikurangi sehingga model dapat lebih fokus pada fitur-fitur yang paling relevan. Proses seleksi fitur menggunakan metode IG pada penelitian ini ditunjukkan pada Gambar 2.

3.4. Pembagian Jumlah Data Latih dan Data Uji

Tahap pembagian data merupakan salah satu langkah penting dalam proses pembuatan model *machine learning*. Penentuan proporsi yang tepat antara data pelatihan dan data pengujian diperlukan untuk memastikan model yang dihasilkan memiliki performa yang baik. Pada penelitian ini, *dataset* dibagi menjadi dua bagian utama yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model agar dapat mempelajari pola-pola yang terdapat dalam dataset, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model terhadap data yang belum pernah digunakan sebelumnya.

Pada penelitian ini digunakan metode pembagian data secara acak dengan perbandingan 80:20, dimana 80% data digunakan sebagai data pelatihan dan 20% data digunakan sebagai data pengujian. Pembagian data dengan proporsi tersebut bertujuan agar model dapat mempelajari pola dari sebagian besar data yang tersedia, sekaligus tetap menyediakan data yang cukup untuk menguji kemampuan model dalam melakukan generalisasi terhadap data baru.

3.5. Model Machine Learning

Machine learning merupakan metode yang memungkinkan sistem komputer untuk mempelajari pola dari data sehingga dapat melakukan prediksi atau klasifikasi terhadap data baru. Dalam penelitian ini, *machine learning* digunakan untuk melakukan proses klasifikasi terhadap data pasien yang terindikasi penyakit cacar monyet. Adapun algoritma *machine learning* yang digunakan dalam penelitian ini yaitu SVM, RF, dan XGBoost. Ketiga algoritma tersebut dipilih karena memiliki kemampuan yang baik dalam menangani permasalahan klasifikasi serta sering digunakan dalam berbagai penelitian di bidang *data mining* dan *machine learning*.

3.6. Hyperparameter Tuning

Hyperparameter tuning merupakan proses penyesuaian parameter-parameter yang ditentukan sebelum proses pelatihan model dilakukan. Parameter tersebut tidak dipelajari langsung oleh model dari data, tetapi harus ditentukan terlebih dahulu oleh peneliti agar model dapat bekerja secara optimal. Dengan melakukan *hyperparameter tuning*, model diharapkan mampu menghasilkan performa yang lebih baik dalam hal akurasi, stabilitas, dan kemampuan generalisasi terhadap data baru.

Pada penelitian ini, proses *hyperparameter tuning* dilakukan menggunakan metode *GridSearchCV* pada model *machine learning* yang digunakan. Metode ini bekerja dengan cara menguji berbagai kombinasi nilai parameter yang telah ditentukan sebelumnya untuk menemukan konfigurasi parameter terbaik. Melalui proses tersebut, model yang dihasilkan diharapkan memiliki performa yang lebih optimal dalam melakukan klasifikasi terhadap data pasien yang terindikasi penyakit cacar monyet.

4. HASIL DAN PEMBAHASAN

4.1. Persiapan Data

Pada tahap ini dilakukan proses persiapan data, yaitu *dataset* yang disimpan dalam file CSV diimpor ke dalam *Visual Studio Code*. File CSV tersebut kemudian dibaca menggunakan bahasa pemrograman *Python*, sehingga data siap untuk diproses lebih lanjut. Setelah data disiapkan, kolom *Patient_ID* tidak digunakan dalam proses klasifikasi karena kolom tersebut hanya berfungsi sebagai pengenal unik setiap data pasien dan tidak memiliki nilai informasi yang relevan dalam mempengaruhi hasil klasifikasi. Data yang digunakan pada penelitian ini merupakan *dataset* pasien yang berkaitan dengan penyakit cacar monyet yang berisi berbagai fitur gejala serta kondisi kesehatan pasien.

4.2. Data Cleaning

Pada tahap *data cleaning*, dilakukan proses pembersihan data untuk memastikan kualitas data yang akan digunakan pada tahap analisis selanjutnya. Proses ini meliputi penghapusan atribut yang memiliki nilai kosong serta atribut yang tidak memberikan kontribusi yang signifikan terhadap proses klasifikasi.

Pada *dataset* cacar monyet, ditemukan adanya nilai NaN (*Not a Number*) pada salah satu atribut, yaitu *Systemic Illness*. Atribut tersebut memiliki data bernilai NaN sebanyak 6216 data, sehingga perlu dilakukan proses pembersihan data untuk mengatasi permasalahan tersebut. Proses pembersihan dilakukan dengan menghapus data yang mengandung nilai NaN agar tidak memengaruhi proses analisis dan pelatihan model.

Selain itu, atribut yang tidak memberikan kontribusi signifikan dalam menentukan kelas serta tidak memiliki korelasi dengan atribut lain juga dihapus dari *dataset*. Langkah ini bertujuan untuk mengurangi kompleksitas data serta meningkatkan kualitas data yang akan digunakan dalam proses pemodelan.

Dengan dilakukannya proses *data cleaning*, *dataset* yang digunakan menjadi lebih bersih, relevan, dan siap untuk diproses pada tahap selanjutnya dalam penelitian.

4.3. Data Transformation

Pada tahap transformasi data, beberapa atribut pada *dataset* diubah bentuknya agar dapat digunakan dalam proses analisis dan pemodelan *machine learning*. Transformasi ini dilakukan karena sebagian data masih berbentuk kategorikal sehingga perlu diubah menjadi nilai numerik.

Atribut yang ditransformasikan antara lain *Rectal Pain*, *Sore Throat*, *Penile Oedema*, *Oral Lesions*, *Solitary Lesion*, *Swollen Tonsils*, *HIV Infection*, dan *Sexually Transmitted Infection*. Proses transformasi dilakukan menggunakan teknik *encoding*, yaitu mengubah data kategorikal menjadi data numerik agar dapat diproses oleh algoritma klasifikasi. Proses ini bertujuan untuk mengubah nilai *True* dan *False* menjadi 1 dan 0, sehingga data yang awalnya berupa nilai logika dapat digunakan dalam proses komputasi.

Pada atribut *Systemic Illness*, dilakukan proses perubahan data dari bentuk kategorikal menjadi numerik. Atribut ini terdiri dari tiga jenis gejala, sehingga masing-masing kategori diubah menjadi nilai angka agar dapat digunakan dalam proses pengolahan data. Dalam proses ini, kategori *Fever* diubah menjadi nilai 1, *Swollen Lymph Nodes* menjadi 2, dan *Muscle Aches and Pain* menjadi 3. Dengan demikian, atribut *Systemic Illness* terbagi menjadi tiga kategori yang direpresentasikan dalam bentuk numerik sehingga mempermudah proses analisis dan pemodelan data.

Pada atribut *Monkeypox* sebagai label kelas, dilakukan penggantian nilai dengan mengasumsikan dua kategori kelas, yaitu *Positive* dan *Negative*, yang kemudian diubah menjadi nilai numerik.

4.4. Seleksi Fitur

Pada tahap seleksi fitur, dilakukan proses pemilihan atribut yang paling relevan terhadap variabel target untuk meningkatkan kinerja model klasifikasi. Metode yang digunakan dalam penelitian ini adalah *Information Gain* (IG), yaitu metode yang digunakan untuk mengukur seberapa besar kontribusi suatu atribut dalam memberikan informasi terhadap kelas atau label data. Nilai IG dihitung berdasarkan pengurangan nilai *entropy* setelah data dibagi berdasarkan atribut tertentu, sehingga atribut dengan nilai IG yang lebih tinggi dianggap memiliki pengaruh yang lebih besar dalam proses klasifikasi. Melalui proses ini, setiap atribut pada *dataset* dihitung nilai IG -nya terhadap atribut target, dan hasil perhitungan tersebut digunakan sebagai dasar untuk menentukan atribut yang paling berpengaruh dalam proses klasifikasi. Nilai IG untuk setiap atribut ditunjukkan pada Tabel 1, yang memuat nama atribut beserta nilai IG yang diperoleh.

Tabel 1. Nilai IG pada setiap Aribut

Atribut	Informatio Gain
Systemic Illness	0,024
Rectal Pain	0,006
Sore Throat	0,000
Penile Oedema	0,003
Oral Lesions	0,003
Solitary Lesion	0,000
Swollen Tonsils	0,002
HIV Infection	0,014
Sexually Transmitted Infection	0,010

Berdasarkan hasil perhitungan IG pada Tabel 1, terlihat bahwa terdapat beberapa atribut yang memiliki nilai IG yang sangat kecil atau mendekati 0. Atribut dengan nilai IG yang mendekati nol menunjukkan bahwa atribut tersebut memberikan informasi yang sangat sedikit terhadap proses klasifikasi. Dalam penelitian ini terdapat dua atribut yang diseleksi yaitu *Sore Throat* dan *Solitary Lesion*, karena keduanya memiliki nilai IG sebesar 0 sehingga dianggap tidak memberikan kontribusi dalam membedakan kelas

pada *dataset*. Sementara itu, atribut lainnya tetap digunakan dalam proses pemodelan karena masih memiliki nilai IG yang lebih besar dari nol, yang menunjukkan bahwa atribut tersebut masih memberikan informasi yang relevan terhadap proses klasifikasi.

4.5. Pembagian Data

Pada tahap pembagian data, *dataset* dibagi menjadi dua bagian yaitu data pelatihan dan data pengujian. Pembagian data dalam penelitian ini menggunakan rasio 80:20, di mana 80% data digunakan untuk proses pelatihan model dan 20% data digunakan untuk proses pengujian model. Dari total 18.784 data yang tersedia pada *dataset*, sebanyak 15.027 data digunakan sebagai *data training* untuk melatih model *machine learning*, sedangkan sisanya sebanyak 3.757 data digunakan sebagai *data testing* untuk mengevaluasi kinerja model yang telah dilatih. Pembagian data ini dilakukan agar model dapat belajar dari sebagian besar data yang tersedia sekaligus dapat diuji kemampuannya dalam melakukan klasifikasi pada data yang belum pernah dilihat sebelumnya.

4.6. Hasil dan Pembahasan

Pada penelitian ini dilakukan beberapa tahapan eksperimen untuk mengetahui pengaruh seleksi fitur, serta optimasi parameter terhadap performa model klasifikasi. Metode klasifikasi yang digunakan yaitu SVM, RF, dan XGBoost. Setiap metode diuji melalui beberapa skenario untuk membandingkan kinerja model yang dihasilkan.

Tahap pertama dilakukan dengan membangun model menggunakan ketiga metode tersebut dengan parameter default tanpa melakukan proses pembersihan data sehingga atribut *Systemic Illness* yang bernilai NaN ditransformasikan menjadi nilai kategorikal yaitu 0. Pada tahap ini, *dataset* digunakan secara langsung setelah proses pembagian data tanpa adanya pengaturan parameter pada model *machine learning*. Tujuan dari tahap ini adalah untuk melihat performa awal dari masing-masing algoritma sebelum dilakukan proses pengolahan data lebih lanjut. Hasil pengujian dari ketiga metode tersebut ditunjukkan pada Tabel 2.

Tabel 2. Performa Model Tahap 1

Model	Akurasi	F1-Score
RF	68,9%	78,3%
SVM	70,1%	79,8%
XGBoost	69,8%	79,1%

Berdasarkan hasil pengujian pada tahap pertama, terlihat bahwa performa awal dari ketiga algoritma machine learning yang digunakan masih berada pada tingkat yang relatif moderat karena model dibangun menggunakan parameter *default* dan tanpa dilakukan proses pembersihan data. Dari hasil yang ditunjukkan pada tabel, metode SVM memperoleh nilai akurasi tertinggi yaitu 70,1% dengan *F1-Score* sebesar 79,8%, sehingga menunjukkan kemampuan yang sedikit lebih baik dalam melakukan klasifikasi dibandingkan metode lainnya. Sementara itu, XGBoost menghasilkan akurasi sebesar 69,8% dengan *F1-Score* 79,1%, yang menunjukkan performa yang cukup kompetitif meskipun masih berada di bawah SVM. Adapun RF memperoleh nilai akurasi paling rendah yaitu 68,9% dengan *F1-Score* 78,3%.

Perbedaan performa ini dapat disebabkan oleh kondisi *dataset* yang masih mengandung nilai yang belum diproses secara optimal, seperti atribut *Systemic Illness* yang memiliki nilai NaN dan hanya ditransformasikan menjadi nilai numerik tanpa proses pembersihan data yang lebih mendalam. Oleh karena itu, hasil pada tahap ini digunakan sebagai *baseline* untuk membandingkan peningkatan performa model pada tahap selanjutnya setelah dilakukan proses pengolahan data lebih lanjut.

Tahap kedua dilakukan dengan menggunakan tiga metode yang sama, yaitu SVM, RF, dan XGBoost dengan parameter *default*, namun pada tahap ini dataset terlebih dahulu melalui proses *data cleaning*. Proses pembersihan data bertujuan untuk menghilangkan data yang tidak lengkap atau tidak relevan sehingga kualitas data yang digunakan dalam proses pelatihan model menjadi lebih baik. Hasil pengujian pada tahap ini ditunjukkan pada Tabel 3.

Tabel 3. Performa Model Tahap 2

Model	Akurasi	*F1-Score*
RF	69,1%	79,5%
SVM	70,9%	81,4%
XGBoost	70,6%	80,7%

Berdasarkan hasil pengujian pada tahap kedua, terlihat bahwa penerapan proses pengolahan data memberikan peningkatan performa pada seluruh metode yang digunakan dibandingkan dengan tahap sebelumnya. Proses pembersihan data membantu menghilangkan data yang tidak lengkap atau tidak relevan sehingga kualitas *dataset* menjadi lebih baik dan lebih representatif untuk proses pelatihan model.

Dari hasil pada tabel, metode SVM menunjukkan performa terbaik dengan nilai akurasi sebesar 70,9% dan *F1-Score* sebesar 81,4%, yang menunjukkan kemampuan klasifikasi yang lebih baik dibandingkan metode lainnya. Sementara itu, XGBoost memperoleh akurasi 70,6% dengan *F1-Score* 80,7%, yang juga menunjukkan peningkatan dibandingkan tahap sebelumnya. Adapun RF menghasilkan akurasi 69,1% dan *F1-Score* 79,5%. Meskipun RF masih menjadi yang terendah di antara ketiga metode, RF tetap menunjukkan peningkatan performa setelah dilakukan proses pengolahan data.

Hasil ini menunjukkan bahwa kualitas data memiliki pengaruh yang cukup signifikan terhadap performa model *machine learning*, di mana *dataset* yang telah dibersihkan mampu membantu algoritma dalam mempelajari pola data dengan lebih baik sehingga menghasilkan prediksi yang lebih akurat.

Tahap ketiga dilakukan dengan menerapkan proses pengolahan data serta menambahkan proses seleksi fitur menggunakan metode IG. Seleksi fitur bertujuan untuk memilih atribut yang paling relevan terhadap variabel target sehingga dapat mengurangi kompleksitas data dan meningkatkan performa model klasifikasi. Pada tahap ini, metode yang digunakan tetap sama yaitu SVM, RF, dan XGBoost dengan parameter *default*. Hasil pengujian dari tahap ini ditunjukkan pada Tabel 4.

Tabel 4. Performa Model Tahap 3

Model	Akurasi	*F1-Score*
RF	70,5%	80,7%
SVM	71%	81,4%
XGBoost	70,3%	80,5%

Berdasarkan hasil pengujian pada tahap ketiga, terlihat bahwa seleksi fitur menggunakan metode IG memberikan peningkatan performa pada sebagian besar model dibandingkan tahap sebelumnya. Proses seleksi fitur berperan dalam memilih atribut yang paling relevan terhadap variabel target sehingga dapat mengurangi atribut yang kurang berpengaruh serta menurunkan kompleksitas data yang diproses oleh model. Dari hasil pada tabel, metode SVM menunjukkan performa terbaik dengan akurasi sebesar 71% dan *F1-Score* sebesar 81,4%, yang menunjukkan bahwa metode ini mampu memanfaatkan fitur-fitur terpilih dengan cukup baik dalam proses klasifikasi. Sementara itu, RF mengalami peningkatan performa dengan akurasi 70,5% dan *F1-Score* 80,7%, sedangkan XGBoost memperoleh akurasi 70,3% dengan *F1-Score* 80,5%.

Secara keseluruhan, hasil ini menunjukkan bahwa penerapan seleksi fitur menggunakan IG dapat membantu model dalam memfokuskan pembelajaran pada atribut yang paling informatif, sehingga mampu meningkatkan akurasi serta kualitas klasifikasi yang dihasilkan.

Tahap keempat merupakan tahap lanjutan dengan menerapkan seluruh proses sebelumnya yaitu pengolahan data dan seleksi fitur, kemudian ditambahkan proses optimasi parameter menggunakan GridSearchCV. Optimasi parameter dilakukan untuk menemukan kombinasi parameter terbaik pada masing-masing algoritma sehingga dapat menghasilkan performa model yang lebih optimal. Metode yang digunakan tetap sama yaitu SVM, RF, dan XGBoost dengan parameter model ditentukan berdasarkan hasil pencarian terbaik dari GridSearchCV. Hasil dari tahap ini ditunjukkan pada Tabel 5.

Tabel 5. Performa Model Tahap 4

Model	Akurasi	*F1-Score*
RF	71,2%	81,6%
SVM	70,8%	81,2%
XGBoost	71,3%	81,4%

Berdasarkan hasil pengujian pada Tabel 5, terlihat bahwa penerapan seleksi fitur IG, serta optimasi parameter menggunakan GridSearchCV mampu memberikan peningkatan performa yang lebih optimal dibandingkan tahap-tahap sebelumnya. Dari ketiga metode yang digunakan, XGBoost memperoleh nilai akurasi tertinggi yaitu 71,3% dengan *F1-Score* sebesar 81,4%. RF dengan akurasi 71,2% dan *F1-Score* 81,6% yang menjadi nilai *F1-Score* tertinggi di antara model lainnya. Sementara itu, SVM memperoleh akurasi sebesar 70,8% dengan F1-Score 81,2%.

Hasil ini menunjukkan bahwa proses optimasi parameter mampu meningkatkan kemampuan model dalam melakukan klasifikasi dengan lebih baik karena parameter yang digunakan telah disesuaikan dengan karakteristik *dataset*. Kombinasi antara data cleaning dan seleksi fitur juga berperan dalam meningkatkan kualitas data yang digunakan sehingga model dapat mempelajari pola dengan lebih optimal. Secara keseluruhan, tahap keempat menghasilkan performa terbaik dibandingkan tahap sebelumnya, yang menunjukkan bahwa integrasi antara pengolahan data dan optimasi model sangat berpengaruh dalam meningkatkan kinerja klasifikasi penyakit cacar monyet.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan seleksi fitur menggunakan IG, serta optimasi parameter menggunakan GridSearchCV memberikan pengaruh positif terhadap peningkatan performa model klasifikasi. Proses pengolahan data yang baik mampu meningkatkan kualitas *dataset* sehingga model *machine learning* dapat mempelajari pola data dengan lebih efektif, sementara seleksi fitur membantu mengurangi kompleksitas data dengan memilih atribut yang paling relevan terhadap variabel target. Selain itu, optimasi parameter mampu menghasilkan konfigurasi model yang lebih optimal sehingga meningkatkan kinerja model. Hasil penelitian menunjukkan bahwa performa terbaik diperoleh pada metode XGBoost dengan nilai akurasi sebesar 71,3% dan *F1-Score* sebesar 81,4%. Untuk pada penelitian selanjutnya disarankan untuk mencoba metode seleksi fitur lainnya, menerapkan teknik optimasi parameter yang berbeda, serta menambahkan algoritma klasifikasi lain agar dapat memperoleh performa model yang lebih baik dan hasil penelitian yang lebih komprehensif.

DAFTAR PUSTAKA

[1] O. Mitjà *et al.*, “Monkeypox,” Jan. 07, 2023, *Elsevier B.V.* doi: 10.1016/S0140-6736(22)02075-X.

[2] A. Gessain, E. Nakoune, and Y. Yazdanpanah, “Monkeypox,” *New England Journal of Medicine*, vol. 387, no. 19, pp. 1783–1793, Nov. 2022, doi: 10.1056/NEJMra2208860.

[3] F. Aldi, I. Nozomi, R. B. Sentosa, and A. Junaidi, "Machine Learning to Identify Monkey Pox Disease," *Sinkron*, vol. 8, no. 3, pp. 1335–1347, Jul. 2023, doi: 10.33395/sinkron.v8i3.12524.

[4] M. D. D. Krisna and F. N. Hasan, "Analisa Kinerja Algoritma Random Forest dan XGBoost dalam Klasifikasi Penyakit Cacar Monyet (Monkeypox)," *Journal of Information System Research (JOSH)*, vol. 6, no. 3, pp. 1757–1766, Apr. 2025, doi: 10.47065/josh.v6i3.7167.

[5] W. Anugrah, E. Haerani, Y. Yusra, and L. Oktavia, "Klasifikasi Penyakit Cacar Monyet Menggunakan Metode Support Vector Machine," *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 3, pp. 558–566, May 2024, doi: 10.47065/josyc.v5i3.5149.

[6] F. Nur, M. Ahsan, and W. Harianto, "KOMPARASI TINGKAT AKURASI INFORMATION GAIN DAN GAIN RATIO PADA METODE K-NEAREST NEIGHBOR," 2022.

[7] D. F. Surianto and K. N. F. Syahnur, "Seleksi Fitur Information Gain (IG) pada Klasifikasi Data Opini Saham menggunakan Metode Naive Bayes," 2023.

[8] D. T. Murdiansyah, "Prediksi Stroke Menggunakan Extreme Gradient Boosting," *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 2, p. 419, Sep. 2024, doi: 10.26798/jiko.v8i2.1295.

[9] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo, and O. O. Oluwagbemi, "Ensemble Machine Learning for Monkeypox Transmission Time Series Forecasting," *Applied Sciences (Switzerland)*, vol. 12, no. 23, Dec. 2022, doi: 10.3390/app122312128.

[10] C. Z. V. Junus, T. Tarno, and P. Kartikasari, "KLASIFIKASI MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK DETEKSI AWAL RISIKO DIABETES MELITUS," *Jurnal Gaussian*, vol. 11, no. 3, pp. 386–396, Jan. 2023, doi: 10.14710/j.gauss.11.3.386-396.

[11] L. Bramasta Erlian Susilo *et al.*, "PENERAPAN ALGORITMA DECISION TREE, RECURSIVE ELIMINATION DAN ADASYN PADA DATA KLASIFIKASI PENYAKIT DERMATITIS," 2025.

[12] I. Arrosyid, D. Hardan Gutama, A. Subhan Yazid, and A. Pramuntadi, "IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK SISTEM PREDIKSI PENYAKIT DIABETES MELITUS TIPE 2," 2025