Journal of Artificial Intelligence and Software Engineering
Vol. 6, No. 2, Juni 2026, pp. 130~141
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v6i2.9384

Application of Information Gain Feature Selection and SMOTE in XGBoost Algorithm for Asthma Disease Classification

Fioni Nikmatul Fajar[1], Fitri Insani[2*], Suwanto Sanjaya[3], Iis Afrianty[4]
[1,2,3,4] Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia

Informasi Artikel

Diterima : 12 Juni 2026
Revisi : 22 Juni 2026
Publikasi : 30 Juni 2026

Kata Kunci:

Asma
Information Gain
SMOTE
XGBoost
Klasifikasi

ABSTRAK

Asma merupakan salah satu penyakit kronis pada sistem pernapasan yang prevalensinya terus meningkat dan memerlukan deteksi dini untuk mencegah komplikasi serius. Salah satu tantangan dalam klasifikasi asma menggunakan *machine learning* adalah ketidakseimbangan kelas yang menyebabkan model cenderung memprediksi kelas mayoritas sehingga kemampuan mendeteksi kasus asma menjadi rendah. Penelitian ini mengusulkan penerapan SMOTE dan seleksi fitur *Information Gain* dalam algoritma XGBoost untuk mengatasi permasalahan tersebut. Dataset yang digunakan terdiri dari 2.392 data dengan 28 atribut, di mana tahapan penelitian meliputi *preprocessing*, seleksi fitur menggunakan *Information Gain* yang mengurangi fitur menjadi 22 fitur, penyeimbangan data menggunakan SMOTE, pembagian data dengan rasio 90:10, 80:20, dan 70:30, serta klasifikasi menggunakan XGBoost. Pengujian dilakukan terhadap empat skenario pendekatan untuk membandingkan kontribusi setiap metode yang diterapkan. Evaluasi dilakukan menggunakan data uji seimbang dan data uji asli dengan metrik akurasi, presisi, *recall*, dan F1-*score*. Hasil penelitian menunjukkan bahwa skenario terbaik diperoleh pada kombinasi *Information Gain* + SMOTE + XGBoost dengan rasio 90:10 pada data uji seimbang, menghasilkan akurasi 75%, presisi 87,5%, *recall* 58,33%, dan F1-*score* 70%. Hasil tersebut menunjukkan bahwa kombinasi seleksi fitur dan penyeimbangan data mampu meningkatkan kemampuan model dalam mendeteksi penyakit asma.

ABSTRACT

Asthma is a chronic respiratory disease whose prevalence continues to increase and requires early detection to prevent serious complications. One of the main challenges in asthma classification using machine learning is class imbalance, which causes models to favor the majority class and reduces their ability to accurately detect asthma cases. This study proposes the integration of SMOTE and Information Gain feature selection with the XGBoost algorithm to address this issue. The dataset used consists of 2,392 records with 28 attributes. The research stages include data preprocessing, feature selection using Information Gain that reduced the number of features to 22, data balancing using SMOTE, data splitting with ratios of 90:10, 80:20, and 70:30, and classification using XGBoost. Four experimental scenarios were conducted to evaluate the contribution of each method. Model performance was assessed using balanced and original test datasets based on accuracy, precision, recall, and F1-score metrics. The results show that the best performance was achieved by the Information Gain + SMOTE + XGBoost combination with a 90:10 data split on the balanced test dataset, yielding an accuracy of 75%, precision of 87.5%, recall of 58.33%, and an F1-score of 70%. These findings indicate that the combination of feature selection and data balancing techniques can improve the model's ability to detect asthma cases.

Penulis Koresponden

Email : fitri.insani@uin-suska.ac.id

Cara sitasi IEEE:

[1] F. N. Fajar, F. Insani, S. Sanjaya, I. Afrianty, "Application of Information Gain Feature Selection and SMOTE in XGBoost Algorithm for Asthma Disease Classification," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 6, no. 2, p. 130-141, Juni 2026. doi:10.30811/jaise.v6i2.9384

1. PENDAHULUAN

Asma merupakan penyakit peradangan kronis pada saluran pernapasan, ditandai dengan penyempitan saluran napas dan peningkatan sensitivitas, dengan gejala seperti batuk, sesak napas, mengi, dan dada yang terasa berat. Kondisi ini juga disertai dengan masalah terkait seperti rinitis, sinusitis, dan dermatitis atopik. Selama beberapa dekade terakhir, prevalensi asma telah meningkat secara signifikan. Lebih dari 300 juta orang mengalaminya, dan diperkirakan akan mencapai 400 juta orang [1], [2]. Penyakit ini ditandai dengan hiperreaktivitas bronkus, penyempitan saluran pernapasan yang tidak stabil, sehingga dapat mengganggu aktivitas sehari-hari dan menimbulkan beban ekonomi yang besar bagi sistem pelayanan kesehatan [3]. Data *World Health Organization* (WHO) menunjukkan bahwa angka kejadian asma di dunia mencapai 262 juta pada tahun 2019, yang menyebabkan 455.000 kematian [4]. Diagnosis yang cepat dan akurat sangat penting dalam pengelolaan klinis untuk mencegah komplikasi yang lebih serius dan diagnosis asma yang akurat. Meskipun demikian, masih ada sejumlah masalah dalam proses diagnosis, termasuk kebutuhan akan pengamatan manual oleh dokter, kurangnya data medis, dan sistem pendukung keputusan berbasis data yang tidak memadai. Selain itu, sebelum digunakan dalam proses klasifikasi, data rekam medis harus ditangani dengan hati-hati karena sering kali formatnya tidak terstruktur, seperti *missing value*, duplikat, dan kombinasi data numerik dan kategorikal [5].

Di era kemajuan teknologi informasi saat ini, penggunaan *machine learning* dan *data mining* semakin banyak digunakan dalam industri kesehatan, terutama dalam pemrosesan data klinis untuk membantu pengambilan keputusan medis, termasuk dalam diagnosis asma [6], [2]. Beberapa penelitian terdahulu telah melakukan klasifikasi penyakit asma menggunakan *machine leaning*, seperti penelitian oleh Lee *et al.* mengenai klasifikasi asma yang mengkombinasikan *Enhanced* GAN dan algoritma XGBoost yang mendapatkan hasil akurasi sebesar 94,03% [2]. Selanjutnya penelitian oleh Dullah *et al.* mengklasifikasikan asma menggunakan *Adaptive Boost* dengan teknik *sampling* SVM-SMOTE menghasilkan akurasi sebesar 98,60% [7]. Penelitian lainnya oleh Sodiq *et al.* yang melakukan klasifikasi asma pada anak-anak dengan membandingkan LR, RF, XGBoost, dan SMOTE untuk *oversampling*. Hasil penelitian tersebut yaitu LR memberikan keseimbangan paling baik antara *precision* dan *recall*, namun XGBoost memilki *recall* tertinggi sebesar 0,66 sehingga mampu mendeteksi lebih banyak kasus asma. Sedangkan RF memperoleh ROC-AUC tertinggi sebesar 0,83 [8]. Selanjutnya penelitian Sajiwo *et al.* juga menggunakan XGBoost untuk klasifikasi ISPU dengan mengkombinasikan dengan teknik *imbalanced* data SMOTE. Hasil dari penelitian tersebut mendapatkan akurasi sangat tinggi sebesar 99,63% [9].

Banyak penelitian yang telah menunjukkan bahwa salah satu algoritma *machine learning* yang paling efektif adalah *Extreme Gradient Boosting* (XGBoost), yaitu metode *ensemble learning* yang mampu menghasilkan model dengan tingkat akurasi yang baik. Algoritma ini dapat menangani berbagai jenis data, termasuk *dataset* berukuran besar dan kompleks. Selain itu, XGBoost dapat mengurangi kemungkinan *overfitting* yang membuat prediksi lebih akurat dan tidak mudah terpengaruh oleh *noise* pada data [10]. Beberapa penelitian terbaru menunjukkan efektivitas penggunaan XGBoost dalam klasifikasi penyakit. Seperti penelitian yang dilakukan oleh Yang *et al.* untuk memprediksi penyakit jantung membandingkan peforma algoritma XGBoost dengan lima algoritma lainnya yaitu RF, KNN, DT, NB, LR. Hasil penelitian menunjukkan bahwa algoritma XGBoost memberikan peforma terbaik dibandingkan kelima algoritma lainnya dengan nilai akurasi 93,44%, presisi 92,66%, *recall* 97,16%, F1-*score* 94,86%, dan AUC 92,44% [11]. Berikutnya penelitian Opitasari *et al.* mengenai klasifikasi diagnosis penyakit kanker serviks menggunakan XGBoost memperoleh hasil nilai akurasi sebesar 86%, presisi 100%, *recall* 82%, dan F1-*score* 90% [12]. Penelitian Sajiwo *et al.* juga menggunakan algoritma XGBoost dan SMOTE untuk klasifikasi ISPU yang memperoleh akurasi yang sangat tinggi sebesar 99,63% [13]. Meskipun algoritma XGBoost terbukti efektif dalam klasifikasi, ketidakseimbangan kelas pada dataset menjadi tantangan tersendiri dalam pengembangan model yang andal dan mampu mengenali kasus positif asma secara lebih baik.

Dalam konteks kesehatan, ketidakseimbangan kelas merupakan masalah yang sering terjadi. Ketika ini terjadi, algoritma klasifikasi cenderung fokus pada kelas mayoritas yang mengakibatkan kemampuan model

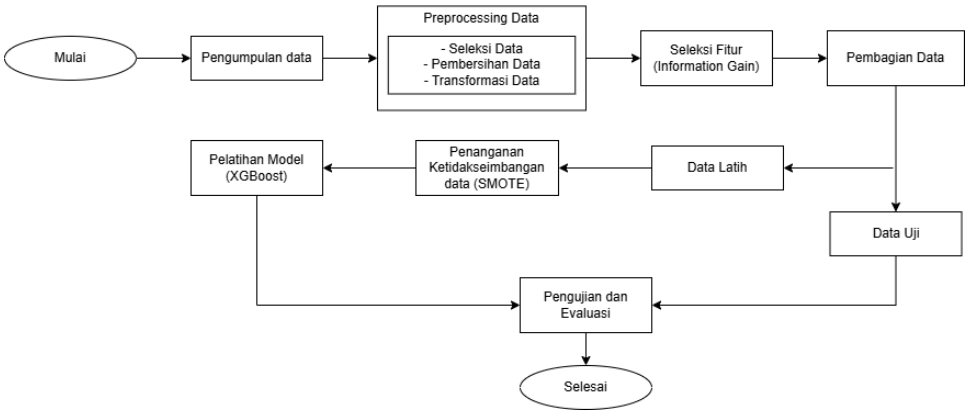
mengidentifikasi pasien positif menjadi rendah dan berpotensi menimbulkan risiko dalam bidang medis [14]. Mengatasi permasalahan tersebut, teknik *oversampling* SMOTE sering digunakan untuk mengatasi masalah ini. SMOTE merupakan metode untuk menambah jumlah data pada kelas minoritas dengan membuat sampel sintetis berdasarkan tetangga terdekat menggunakan jarak *Euclidean*, sehingga distribusi data menjadi lebih seimbang dengan kelas mayoritas [15], [16]. Beberapa penelitian terdahulu telah melakukan penelitian dengan teknik SMOTE yang menunjukkan peforma baik, seperti penelitian Sajiwo *et al.* yang menggunakan XGBoost dan SMOTE memperoleh akurasi tinggi yaitu 99,63% [9]. Penelitian lain oleh Sijabat *et al.* menggunakan SMOTE dan membandingkan empat algoritma *machine learning* lainnya, yaitu LR, RF, *LightGBM*, dan XGBoost untuk klasifikasi penyakit jantung yang mampu menghasilkan akurasi di atas 0,83 dan nilai AUC-ROC di atas 0,90 [17]. Selain itu, penelitian oleh Illawati *et al.* pada klasifikasi penyakit *Monkeypox* yang mengimplementasikan XGBoost dan SMOTE dan mampu mendapatkan hasil *recall* mencapai 0,93 [10].

Dalam proses analisis data, pemilihan fitur memiliki peran sangat penting. Tahapan ini bertujuan memilih dan mengekstrak fitur yang paling relevan dari data awal, sehingga dapat mengurangi jumlah dimensi pada *dataset* yang besar serta menghindari permasalahan terkait kompleksitas data yang dapat membantu meningkatkan kinerja metode klasifikasi [18]. Dengan memilih fitur yang tepat, model akan menjadi lebih efisien dan akurat [13]. *Information Gain* adalah salah satu metode yang sering digunakan dalam *machine learning* untuk memilih atribut. Metode ini menilai dan memberi peringkat setiap atribut serta menghapus atribut yang tidak memenuhi kriteria [19]. Penerapan *Information Gain* dalam meningkatkan akurasi model telah dibuktikan melalui berbagai penelitian, seperti penelitian oleh Ulinnuha *et al.* melakukan klasifikasi menggunakan *Naive Bayes* dengan seleksi fitur *Information Gain* menghasilkan akurasi terbaik 98,36%, dibandingkan tanpa seleksi fitur dengan akurasi 91,23% [20]. Penelitian oleh Pardede *et al.* menggunakan lima metode seleksi fitur, yaitu *Chi-Square* (CS), *Information Gain* (IG), *Genetic Algorithm* (GA), *Particle Swarm Optimization* (PSO), dan *Least Absolute Shrinkage and Selection Operator* (LASSO), dan tiga metode pengklasifikasi yaitu *Naïve Bayes*, XGBoost dan RF *Classifier*. Hasilnya menunjukkan *Chi-Square* dan *Information Gain* sebagai metode seleksi fitur terbaik, terutama ketika dikombinasikan dengan XGBoost yang meningkatkan akurasi hingga 1,7%, *recall* 2,3%, dan waktu pelatihan sekitar 23% [19]. Penelitian Devian *et al.* yang memprediksi penyakit diabetes mengguanakan KNN dengan *Information Gain,* dengan hasil model tanpa seleksi fitur mendapatkan akurasi 69,11% (rasio 90:10), sedangkan dengan *Information Gain* meningkat menjadi 70,96%, dengan K=17 mencapai akurasi tertinggi 72,93% [21].

Penelitian Sarman *et.al* menunjukkan bahwa klasifikasi penyakit asma dapat dilakukan dengan algoritma dasar seperti *Naïve Bayes* dan SVM, namun penelitian tersebut masih terbatas pada perbandingan model tanpa penguatan melalui seleksi fitur maupun penanganan ketidakseimbangan data [5]. Di sisi lain, penelitian Dullah *et al*. telah menggunakan pendekatan yang lebih optimal dengan *Adaptive Boosting* dan SVM-SMOTE untuk menyeimbangkan data, tetapi masih belum mengeksplorasi kombinasi *Information Gain* dan XGBoost [7]. Penelitian ini menerapkan SMOTE secara ketat hanya pada data latih setelah pembagian data, sehingga evaluasi mencerminkan performa model pada data yang belum pernah dilihat model. Model juga dievaluasi pada dua kondisi data uji, yaitu data uji seimbang dan data uji asli, untuk memberi gambaran performa yang lebih transparan. Pengujian dilakukan pada tiga rasio pembagian data untuk melihat konsistensi performa model. Visualisasi t-SNE juga digunakan untuk menunjukkan tumpang tindih kelas asma dan tidak asma sebagai dasar penerapan SMOTE. Oleh karena itu, masih terdapat celah penelitian untuk menguji apakah *Information Gain*, SMOTE, dan XGBoost dapat menghasilkan performa klasifikasi asma yang lebih baik, khususnya pada metrik F1-*score* dan kemampuan deteksi kelas minoritas. Berdasarkan hal tersebut, penelitian ini mengusulkan pendekatan klasifikasi penyakit asma dengan menggabungkan algoritma XGBoost, *Information Gain*, dan SMOTE untuk mencari kombinasi yang paling optimal dalam menghasilkan model klasifikasi penyakit asma yang akurat dan andal, sehingga dapat mendukung proses pengambilan keputusan medis yang lebih cepat dan tepat.

2. METODE

Penelitian ini memiliki beberapa tahapan untuk klasifikasi penyakit asma menggunakan seleksi fitur *Information Gain*, teknik penyeimbang data SMOTE, dan algoritma XGBoost. Untuk mencapai hasil evaluasi kinerja model, proses penelitian dilakukan melalui beberapa tahapan utama yang dilakukan secara sistematis. Adapun tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian Klasifikasi Asma

2.1 Pengumpulan Data

Data yang digunakan pada penelitian ini berupa data sekunder yang diperoleh dari situs *Kaggle* tahun 2024 yang bisa diakses melalui https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset. *Asthma Disease Dataset* berisi informasi kesehatan 2.392 pasien dengan 1 kelas target dan 28 Fitur yang mencakup informasi demografis, gaya hidup, riwayat medis, hasil periksaan klinis, serta label diagnosis asma. Fitur pada *dataset* dapat dilihat pada Tabel 1.

Tabel 1. Fitur pada *dataset*

No	Fitur	Deskripsi
1	*PatientID*	Id Pasien
2	*Age*	Usia
3	*Gender*	Jenis kelamin
4	*Ethnicity*	Etnis
5	*EducationLevel*	Tingkat pendidikan
6	*BMI*	Indeks massa tubuh
7	*Smoking*	Status merokok
8	*PhysicalActivity*	Aktivitas fisik mingguan dalam beberapa jam
…	…	…
29	*DoctorInCharge*	Dokter yang bertanggungjawab

2.2 Preprocessing Data

2.2.1 Seleksi Data

Tahap seleksi data dikenal sebagai proses pemilihan fitur yang relevan untuk digunakan dalam proses penelitian klasifikasi penyakit asma. Tahapan ini merupakan proses penghapusan fitur yang tidak memiliki kontribusi prediktif terhadap hasil klasifikasi yang dapat menyebabkan model hanya menghafal data tertentu dan sulit diterapkan pada data pasien lain.

2.2.2 Pembersihan Data

Tahap ini membersihkan data dan menata ulang yang bertujuan agar data sesuai dengan kebutuhan pemodelan yang akan dilakukan. Tahap ini meliputi pengecekan *missing value*, data duplikat, dan *outlier* pada atribut numerik. Hal ini dilakukan untuk menghindari bias akibat data yang hilang dan memastikan bahwa semua atribut siap digunakan dalam pelatihan model.

2.2.3 Transformasi Data

Transformasi data yaitu proses menggabungkan, mengubah atau menata ulang data kedalam format tertentu agar sesuai dengan pemodelan yang akan dilakukan. Normalisasi menjadi langkah untuk menyelaraskan nilai-nilai fitur dari data agar berada dalam batas tertentu [22]. Pada penelitian ini dilakukan normalisasi data menggunakan metode *MinMax* pada data kontinu. *MinMax* dihitung dengan rumus pada Persamaan (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Pada Persamaan (1), *x* merupakan nilai data asli yang akan dinormalisasi. *min (x)* adalah nilai minimun dari dataset, sedangkan *max (x)* merupakan nilai maksimum dari *dataset* [16].

2.3 Seleksi Fitur Information Gain

Seleksi Fitur dikenal sebagai tahap untuk menentukan variabel yang paling penting guna meningkatkan kinerja model dan mempercepat proses pelatihan dengan menganalisis hubungan antara variabel *input* dan *output*. Pada proses ini, variabel dengan korelasi rendah akan dihilangkan sehingga hanya tersisa fitur yang berkualitas [23]. Penelitian ini menerapkan metode seleksi fitur *Information Gain*, yaitu metode yang berfungsi untuk menyaring fitur-fitur penting dan mengurangi jumlah fitur pada *dataset* yang diawali dengan proses perhitungan entropi sebelum dan sesudah data dipisahkan [21], [16]. Entropi dihitung dengan rumus pada Persamaan (3).

$$Entropy\ (S) = -\sum_{i=1}^{n} p_i \log 2\, p_i \tag{3}$$

Pada Persamaan (3), *pi* adalah jumlah sampel pada kelas ke-i [24]. Sedangkan untuk menghitung *Information Gain* dapat dilihat pada rumus Persamaan (4).

$$Gain(S, A) = Entropy(s) - \Sigma_{values\,(A)} \frac{|Sv|}{|S|}\ Entropy\ (Sv) \tag{4}$$

Pada Persamaan (4), *A* menyatakan atribut pada data, sedangkan *values(A)* merupakan kumpulan nilai yang dimiliki atribut *A*, $|Sv|$ menunjukkan jumlah sampel nilai *v*, dan $|S|$ adalah total seluruh sampel data. Sementara itu *Entropy* (Sv) merupakan nilai entropi dari kumpulan sampel yang memiliki nilai v [24].

2.4 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE merupakan teknik penyeimbangan data yang bertujuan menambah jumlah sampel pada kelas dengan data paling sedikit melalui pembuatan data sintetis. Jumlah data pada kelas minoritas ditingkatkan hingga mendekati atau menyamai kelas dengan jumlah data terbanyak [7]. Adapun rumus pencarian data sintetis SMOTE pada Persamaan (2).

$$Xsyn = Xi + (Xknn - Xi) \times \delta \tag{2}$$

Pada Persamaan (2), $Xsyn$ merupakan data sintetis yang dihasilkan oleh SMOTE. Data sintetis ini dibentuk berdasarkan Xi, yaitu data pada kelas minoritas yang dipilih sebagai sampel awal. Kemudian, $Xknn$ merupakan data tetangga terdekat dari Xi. δ adalah nilai acak antara 0 dan 1 yang digunakan dalam proses pembentukan data sintetis baru [25].

2.5 XGBoost (eXtreme Gradient Boosting)

XGBoost merupakan metode *machine learning* berbasis teknik *boosting* yang digunakan untuk proses klasifikasi dan prediksi dengan memanfaatkan beberapa pohon keputusan yang saling bergantung untuk memperbaiki kelemahan pada pohon sebelumnya. Dalam membangun pohon klasifikasi yang optimal, XGBoost menyusaikan bobot pada setiap pohon yang kemudian seluruh bobot dijumlahkan dan dimasukkan ke dalam fungsi logistik pada tahap prediksi [26], [27].

$$\hat{y}_i = \sum_{m=1}^{m} f_m\ (x_i), f_m\ \epsilon\ F \tag{5}$$

Pada Persamaan (5), yi menunjukkan nilai prediksi untuk data ke-i, sedangkan x_i merupakan fitur yang digunakan dalam proses prediksi. Simbol $\sum_{m=1}^{m}$ menyatakan proses penjumlahan dari semua pohon keputusan, dan f_m merepresentasikan setiap pohon keputusan yang bekerja secara mandiri [13].

Persamaan (6) digunakan untuk menghitung kesalahan prediksi sekaligus *regularization term* yang berfungsi mengontrol kompleksitas model agar dapat mengurangi risiko *overfitting*. Fungsi $\sum l(yi\hat{},yi)$ merupakan *loss function* untuk mengukur selisih antara hasil prediksi dan nilai sebenarnya, sedangkan $\sum \Omega(fi)$ adalah *regularization term* untuk mengatur kompleksitas setiap pohon keputusan [13].

$$L(fi\ ) = \sum l(\hat{y}i\ , yi\ ) + \sum \Omega(f_i) \tag{6}$$

Persamaan (7) digunakan untuk memperbarui nilai prediksi pada setiap iterasi proses pelatihan model [13]. Pada persamaan tersebut $\hat{y}_i^{(t)}$ menunjukkan hasil prediksi pada iterasi ke-*t* untuk data ke-*i*, sedangkan $\hat{y}i^{(t-1)}$

merupakan hasil prediksi pada iterasi sebelumnya. Sementara $fi_{(xi)}$ adalah fungsi yang dihasilkan pada iterasi ke-*t* untuk memperbaiki prediksi data ke-*i* [13].

$$\hat{y}i\ (t) = \hat{y}i^{(t-1)} + fi_{(xi)} \tag{7}$$

Persamaan (8) digunakan untuk menghitung *regularization term* pada keseluruhan model guna mengontrol kompleksitas dan mengurangi risiko *overfitting*. γ merupakan parameter regularisasi yang menentukan besar penalti berdasarkan jumlah *leaf node*, sedangkan *T* menyatakan jumlah *leaf node* pada model. Simbol λ adalah parameter regularisasi yang digunakan untuk mengendalikan kompleksitas model, *w* menunjukkan bobot pada setiap *leaf node*, dan *j* merupakan indeks atau penanda untuk masing-masing *leaf node* [13].

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

Persamaan (9) digunakan untuk melakukan optimasi fungsi *loss* menggunakan pendekatan Taylor orde kedua pada proses *boosting*. $L(t)$ menyatakan fungsi *loss* pada iterasi *bossting* ke-*t*, sedangkan $f_t(xi)$ merupakan *output* model baru pada iterasi ke-*t* terhadap *input xi*. Simbol g_i menunjukkan gradien pertama dari fungsi *loss*, sementara h_i adalah gradien kedua atau hessian dari fungsi *loss*. Adapun $\Omega(f_t)$ ialah fungsi regularisasi yang digunakan untuk mengontrol kompleksitas pohon keputusan dalam model [13].

$$L(t) \approx \sum_{i=1}^{n}[(g_i f_t(xi) + \frac{1}{2}h_i f_t(xi)^2\ ] + \Omega(f_t) \tag{9}$$

Persamaan (10) untuk mengoptimalkan bobot pada setiap *leaf node* menggunakan gradien dan hessian dalam proses pembentukan model XGBoost [13].

$$L(f_i) \approx \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T \tag{10}$$

Persamaan (11) digunakan untuk menghitung nilai *gain* dalam menentukan split terbaik saat membangun pohon keputusan. G_L dan $G_R$menyatakan jumlah gradien pada *node* kiri dan *node* kanan, sedangkan $H_L$dan $H_R$menunjukkan jumlah nilai hessian pada masing-masing *node*. Sementara itu, γ merupakan parameter regularisasi yang digunakan untuk mengontrol minimum pengurangan *loss* agar proses split dapat dilakukan [13].

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{11}$$

Persamaan (12) untuk menentukan bobot optimal pada setiap *leaf node* dalam pohon keputusan. Simbol $w*_j$ menyatakan bobot optimal pada *leaf node* ke-j, sedangkan $G_j$merupakan jumlah gradien pada *leaf node* tersebut dan $H_j$adalah jumlah nilai hessian pada *leaf node* ke-j. Sementara itu, λ merupakan parameter regularisasi yang digunakan untuk mengontrol kompleksitas model [13].

$$w*_j = -\frac{G_j}{H_j+\lambda} \tag{12}$$

2.6 Pengujian dan Evaluasi

Evaluasi kinerja model klasifikasi dilakukan menggunakan *confusion matrix* untuk menganalisis hasil pengujian model. Empat istilah utama digunakan dalam *confusion matrix* untuk menunjukkan hasil dari proses klasifikasi : *True Positif* (TP) menunjukkan jumlah data positif yang berhasil diidentifikasi dengan benar oleh sistem; *True Negatif* (TN) menunjukkan jumlah data negatif yang berhasil diidentifikasi dengan benar oleh sistem; *False Positif* (FP) menunjukkan jumlah data positif yang salah diklasifikasikan oleh sistem; dan *False Negatif* (FN) menunjukkan jumlah data negatif yang salah diklasifikasikan oleh sistem [16]. Peforma algoritma diukur berdasarkan :

a) Akurasi

$$Akurasi = \frac{TN+TP}{TP+TN+FP+FN} \tag{13}$$

b) Presisi

$$Presisi = \frac{TP}{TP + FP} \tag{14}$$

c) Recall

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

d) F1 score

$$F1\ Score = 2\ x \frac{Presisi * Recall}{Presisi + Recall} \tag{16}$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Penelitian ini menggunakan data dari *platform Kaggle* tahun 2024 yang berjudul *Asthma Disease Dataset* yang terdiri dari 2.392 data. Data tersebut berisi jumlah data pasien penderita asma sebanyak 124 data, sedangkan jumlah data pasien tidak asma sebanyak 2.268 data. Dataset awal dapat dilihat pada Tabel 2.

Tabel 2. Dataset awal Penyakit Asma

PatientID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	...	DoctorInCharge
5034	63	0	1	0	15, 84874439	0	...	Dr_Confid
5035	26	1	2	2	22, 75704209	0	...	Dr_Confid
5036	57	0	2	1	18, 39539647	0	...	Dr_Confid
5037	40	1	2	1	38, 51527789	0	...	Dr_Confid
…	...	...	...	...	...	...	...	…
7425	26	1	0	0	28, 12302120	1	...	Dr_Confid

3.2 Preprocessing Data

Tahap *preprocessing* diawali dengan seleksi data melalui penghapusan dua fitur yaitu *PatientID* dan *DoctorInCharge*, karena kedua atribut tersebut hanya bersifat administratif dan identifikasi sehingga tidak memiliki kontribusi prediktif terhadap klasifikasi penyakit. Selain itu, dilakukan pembersihan data dengan pengecekkan terlebih dahulu *missing value*, data duplikat, dan nilai *outlier*. Hasil dari pengecekkan tersebut ialah tidak adanya data yg bermasalah. Selanjutnya dilakukan tahap normalisasi *MinMax* untuk penskalaan fitur kontinu. Fitur-fitur kontinu tersebut dapat dilihat pada Tabel 3 berikut.

Tabel 3. Daftar Fitur Kontinu

Age	BMI	Physical Activity	Diet Quality	Sleep Quality	Pollution Exposure	Pollen Exposure	Dust Exposure	Lung Function FEV1	Lung Function FVC
63	15,848744	8,944483	5,488695	8,701003	7,388481	2,855578	9,743394	1,369051	4,941206
26	22,757042	5,897329	6,341014	5,153966	1,969838	7,457665	6,584631	2,197767	1,702393
57	18,395396	6,739367	9,196237	6,840647	1,460593	1,448187	5,445799	1,698011	5,022553
40	38,515278	1,404503	5,826532	4,253036	5,819053	7,571845	3,965316	3,032037	2,300159
…	...	...	...	...	...	...	...	...	...
26	28,123021	1,613138	7,412878	8,512253	3,231709	3,874028	5,064317	2,280613	2,453284

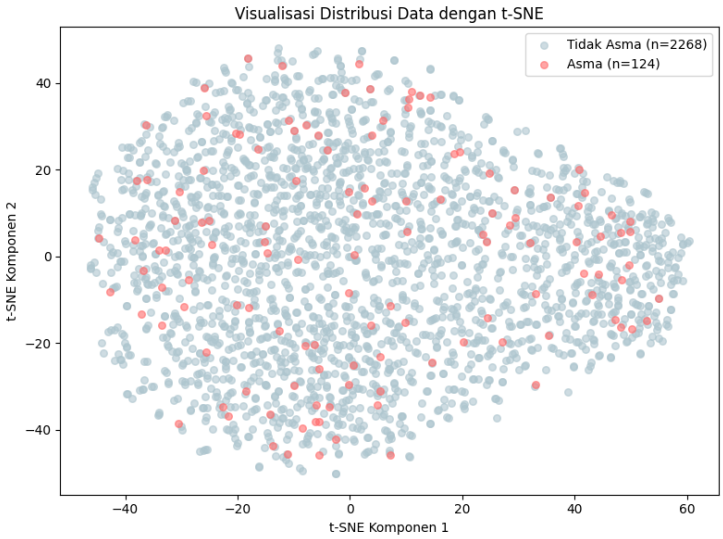
Hasil dari *preprocessing* yang telah dilakukan dapat dilihat pada Tabel 4 berikut ini.

Tabel 4. Hasil *Preprocessing* Data

Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	PhysicalActivity	...	Diagnosis
0.783784	0	1	0	0.032738	0	0,089324	...	0
0.283784	1	2	2	0.309582	0	0,589909	...	0
0.702703	0	2	1	0.134793	0	0,674163	...	0
0.472973	1	2	1	0.941078	0	0,140359	...	0
…	...	...	...	...	...	...	...	...
0.283784	1	0	0	0.524618	1	0,161235	...	0

3.3 Visualisasi Distribusi Data

Sebelum penerapan SMOTE, dilakukan visualisasi distribusi data menggunakan t-SNE (*t-Distributed Stochastic Neighbor Embedding*) untuk melihat sebaran data kelas asma dan tidak asma dalam ruang dua dimensi. Hasil visualisasi t-SNE dapat dilihat pada Gambar 2.



Gambar 2. Visualisasi Persebaran Data

Berdasarkan Gambar 2, terlihat bahwa data kelas asma tersebar secara acak di antara data kelas tidak asma tanpa membentuk batas pisah yang jelas dan menunjukkan bahwa kedua kelas memiliki karakteristik fitur yang tumpang tindih (overlapping) sehingga batas keputusan antar kelas sulit dipelajari oleh model. Maka dari itu, kondisi ini menjadi dasar utama penerapan SMOTE dalam menyeimbangkan distribusi kelas.

3.4 Seleksi Fitur Information Gain

Tahap seleksi fitur ini dilakukan untuk menentukan fitur-fitur yang paling relevan pada proses klasifikasi asma. Proses seleksi fitur diawali dengan perhitungan nilai entropi *dataset* menggunakan Persamaan (3), kemudian dilanjutkan dengan perhitungan *Information Gain* menggunakan Persamaan (4). Dari perhitungan nilai entropi diperoleh nilai entropi sebesar 0,294160. Selanjutnya, nilai *Information Gain* setiap fitur dihitung dengan mencoba seluruh kandidat *threshold* yang merupakan titik tengah antara dua nilai unik berurutan, kemudian dipilih *threshold* yang menghasilkan nilai *Information Gain* tertinggi sebagai nilai akhir fitur. Hasil perhitungan nilai *Information Gain* seluruh fitur dapat dilihat pada Tabel 5 berikut.

Tabel 5. Hasil Perhitungan *Information Gain*

No	Fitur	Information Gain
1	ExerciseInduced	0.002190
2	BMI	0.001578
3	LungFunctionFEV1	0.001532
4	PhysicalActivity	0.001444
5	Age	0.001387
6	DietQuality	0.001361
7	PollenExposure	0.001323
8	DustExposure	0.001295
9	LungFunctionFVC	0.001281
10	EducationLevel	0.001243
11	SleepQuality	0.001230
12	ChestTightness	0.001117
13	PollutionExposure	0.000811
14	Wheezing	0.000543
15	Coughing	0.000423
16	GastroesophagealReflux	0.000355
17	NighttimeSymptoms	0.000344
18	Ethnicity	0.000333
19	Smoking	0.000285
20	HayFever	0.000273
21	ShortnessOfBreath	0.000169
22	PetAllergy	0.000128
23	Eczema	0.000054
24	Gender	0.000007
25	HistoryOfAllergies	0.000003
26	FamilyHistoryAsthma	0.000001

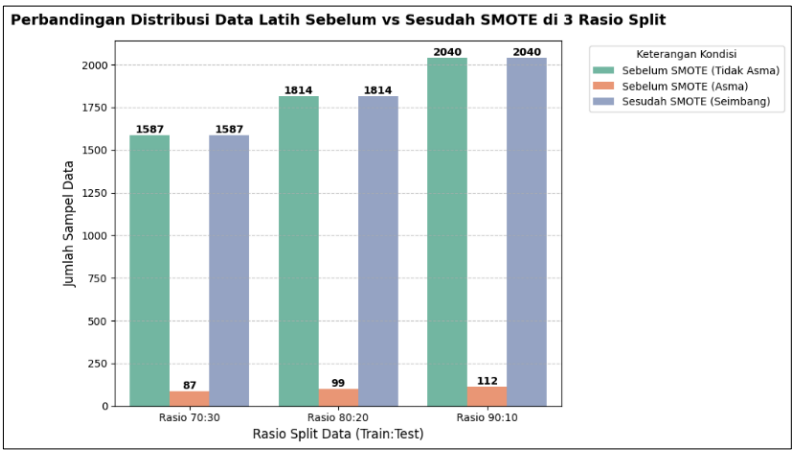
Berdasarkan Tabel 5, seluruh fitur telah dihitung nilai *Information Gain*-nya dan diurutkan dari nilai tertinggi hingga terendah. Untuk menentukan fitur yang relevan, dilakukan percobaan dengan dua nilai

threshold yang berbeda. Percobaan pertama menggunakan *threshold* 0,001 menghasilkan sebanyak 12 fitur terpilih, sedangkan percobaan kedua menggunakan *threshold* 0,0001 menghasilkan sebanyak 22 fitur terpilih.

Hasil percobaan menunjukkan bahwa *threshold* 0,001 menghasilkan nilai metrik evaluasi yang lebih rendah karena beberapa fitur yang masih relevan ikut terbuang sehingga model kekurangan informasi dalam mempelajari pola kelas asma. *Threshold* 0,0001 memberikan performa yang lebih baik karena model memiliki informasi yang lebih lengkap dalam proses pelatihan. Oleh karena itu *threshold* 0,0001 ditetapkan sebagai nilai *threshold* final dengan jumlah fitur terpilih sebanyak 22 fitur.

3.5 SMOTE (Synthetic Minority Over-Sampling Technique)

Pada penelitian ini, *Asthma Disease Dataset* memiliki kelas yang tidak seimbang. Di mana jumlah kelas 0 sebanyak 2.268 data, sedangkan jumlah kelas 1 hanya sebanyak 124 data. Sehingga, dilakukan teknik penyeimbang data SMOTE agar jumlah kelas minoritas menyamai kelas mayoritas. Parameter yang digunakan yaitu *k-neighbors* dengan nilai *default* 5 dan *random_state* = 42. Metode SMOTE diterapkan pada data latih di masing-masing rasio pembagian data. Hasil distribusi kelas setelah SMOTE untuk setiap rasio pembagian data dapat dilihat pada Gambar 3 berikut.



Gambar 3. Hasil Distribusi Kelas Setelah SMOTE pada Setiap Rasio Pembagian Data

3.6 Pengujian dan Evaluasi

Pengujian pada penelitian ini dilakukan menggunakan empat skenario, yaitu XGBoost, *Information Gain* + XGBoost, SMOTE + XGBoost, dan *Information Gain* + SMOTE + XGBoost. Setiap skenario diuji menggunakan tiga rasio pembagian data, yaitu 90 : 10, 80 : 20, dan 70 : 30, untuk mengetahui kombinasi yang paling efektif. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score sesuai Persamaan (13) hingga Persamaan (16). Model terbaik ditentukan berdasarkan nilai F1-score tertinggi, karena metrik ini mampu menunjukkan keseimbangan antara presisi dan recall, terutama pada data yang tidak seimbang. Pelatihan XGBoost diawali dengan percobaan beberapa parameter, yaitu *learning rate* 0,1; 0,01; dan 0,001, *n estimators* 100, 200, 300, 400, dan 500, serta *max depth* 3, 4, 5, 6, dan 7. Berdasarkan hasil percobaan tersebut, parameter *final* yang digunakan adalah *learning_rate* 0,01 untuk mengontrol laju pembelajaran model setiap iterasi, *n_estimators* 300 yang menunjukkan jumlah pohon keputusan yang dibangun, *max depth* 7 yang menentukan kedalaman maksimum setiap pohon.

Pengujian pertama dilakukan menggunakan dataset uji yang telah diseimbangkan dengan metode *RandomUnderSampler*. Skenario ini bertujuan untuk mengukur kemampuan murni algoritma XGBoost secara adil dalam membedakan karakteristik klinis pasien asma dan tidak asma tanpa adanya bias dominasi kelas mayoritas. Pengujian kedua menggunakan data uji asli yang mencerminkan kondisi nyata distribusi kelas yang tidak seimbang untuk melihat performa model pada kondisi sesungguhnya. Hasil dari pengujian tersebut dapat dilihat pada Tabel 6 berikut.

Tabel 6. Hasil Semua Pengujian dengan Data Uji Seimbang

Skenario	Rasio	Akurasi	Presisi	Recall	F1-Score
XGBoost	90 : 10	50%	0%	0%	0%
	80 : 20	50%	0%	0%	0%
	70 : 30	50%	0%	0%	0%
Information Gain + XGBoost	90 : 10	50%	0%	0%	0%
	80 : 20	50%	0%	0%	0%
	70 : 30	50%	0%	0%	0%
SMOTE + XGBoost	90 : 10	70.83%	100%	41.67%	58.82%
	80 : 20	58%	70%	28%	40%
	70 : 30	48.65%	42.86%	8.11%	13.64%

	90 : 10	75%	87.5%	58.33%	70%
Information Gain + SMOTE + XGBoost	80 : 20	58%	70%	28%	40%
	70 : 30	54.05%	63.64%	18.92%	29.17%

Berdasarkan Tabel 6 tersebut, Pada keempat skenario yang telah diuji, hasil tertinggi diperoleh pada skenario keempat yaitu *Information Gain* + SMOTE + XGBoost dengan rasio 90:10 dimana mendapatkan nilai akurasi sebesar 75%, presisi 87,5%, *recall* 58,33%, dan F1-*score* 70%. Hal ini menunjukkan bahwa kombinasi seleksi fitur *Information Gain*, penyeimbangan data SMOTE, dan algoritma XGBoost menghasilkan keseimbangan terbaik antara presisi dan *recall* dibandingkan skenario lainnya. Penambahan seleksi fitur *Information Gain* terbukti memberikan kontribusi positif terhadap performa model dengan mengurangi fitur yang tidak relevan sehingga model dapat fokus mempelajari fitur-fitur dalam membedakan kelas asma dan tidak asma.

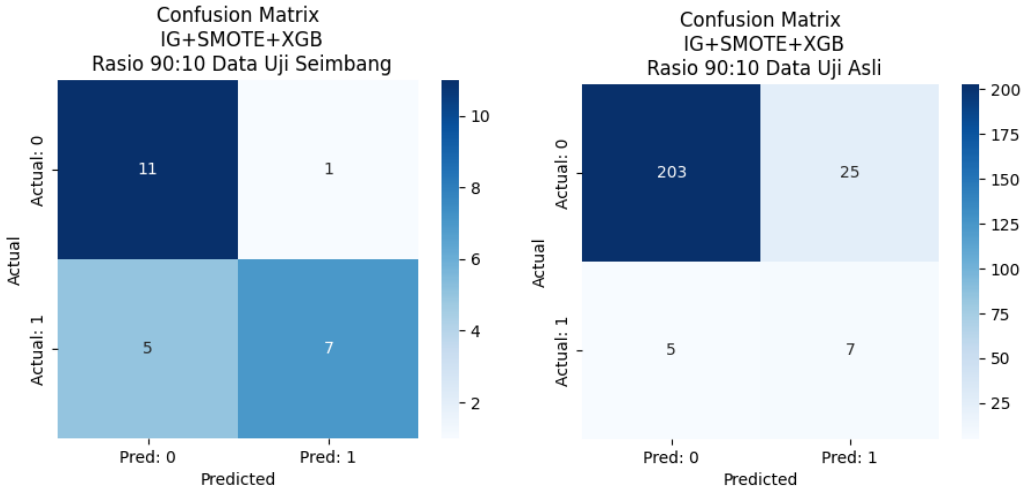
Pengujian kedua dilakukan menggunakan dataset uji asli tanpa penyelarasan jumlah kelas. Hasil dari pengujian tersebut dapat dilihat pada Tabel 7 berikut.

Tabel 7. Hasil Semua Pengujian dengan Data Uji Asli

Skenario	Rasio	Akurasi	Presisi	Recall	F1-Score
	90 : 10	95%	0%	0%	0%
XGBoost	80 : 20	94.78%	0%	0%	0%
	70 : 30	94.85%	0%	0%	0%
	90 : 10	95%	0%	0%	0%
Information Gain + XGBoost	80 : 20	94.78%	0%	0%	0%
	70 : 30	94.85%	0%	0%	0%
	90 : 10	88.75%	20%	41.67%	27.03%
SMOTE + XGBoost	80 : 20	87.27%	14%	28%	18.67%
	70 : 30	85.93%	4.29%	8.11%	5.61%
	90 : 10	87.5%	21.88%	58.33%	31.82%
Information Gain + SMOTE + XGBoost	80 : 20	84.13%	10.77%	28%	15.56%
	70 : 30	82.73%	6.93%	18.92%	10.14%

Berdasarkan Tabel 7, hasil pengujian dengan data uji yang asli tanpa diseimbangkan jumlah data pada kelasnya, mendapatkan hasil F1score tertinggi di skenario keempat pada rasio 90:10 yaitu 31,82%, akurasi sebesar 87,5%, presisi 21,88%, dan *recall* 58,33%. Akurasi yang tinggi tidak selalu berarti model bekerja baik, karena data didominasi oleh kelas negatif. Model cenderung lebih mudah menebak kelas mayoritas sehingga akurasi meningkat, tetapi kemampuan mengenali kelas positif menurun. Hal ini terlihat dari presisi yang rendah dan F1-*score* yang jauh lebih kecil dibandingkan data uji seimbang.

Analisis tambahan dilakukan menggunakan *confusion matrix* untuk mendapatkan gambaran yang lebih baik tentang kinerja model. Analisis ini merinci bagaimana model mengkategorikan setiap sampel ke dalam kelas yang benar dan salah. Hasil *confusion matrix* tersebut dapat dilihat pada Gambar 3.



Gambar 3. Perbandingan *Confusion Matrix*

Berdasarkan Gambar 3, Pada pengujian menggunakan data uji seimbang, model berhasil mengklasifikasikan 11 data kelas negatif (*True Negative*) dan 7 data kelas positif (*True Positive*) dengan benar, serta terdapat 1 data yang salah diklasifikasikan sebagai positif (*False Positive*) dan 5 data yang salah diklasifikasikan sebagai negatif (*False Negative*). Hasil ini menunjukkan bahwa model mampu mengenali kedua kelas secara relatif seimbang karena distribusi data uji telah dibuat proporsional, sehingga nilai presisi, *recall*, dan F1-*score* dapat merepresentasikan kemampuan model secara lebih objektif. Sementara itu, pada

pengujian menggunakan data uji asli yang tidak diseimbangkan, model berhasil mengklasifikasikan 203 data kelas negatif dan 7 data kelas positif dengan benar. Namun, masih terdapat 25 data *False Positive* dan 5 data *False Negative*. Tingginya nilai akurasi pada pengujian ini disebabkan oleh dominasi kelas mayoritas pada data uji yang mencerminkan distribusi asli dataset yang tidak seimbang. Meskipun demikian, model tetap mampu mendeteksi 7 dari 12 kasus asma yang tersedia pada data uji.

Secara keseluruhan, performa model menunjukkan bahwa kombinasi seleksi fitur dan penyeimbangan data memberikan kontribusi positif dalam meningkatkan kemampuan deteksi kelas asma. Penurunan performa pada data uji asli disebabkan oleh beberapa faktor. Pertama, SMOTE hanya diterapkan pada data latih, sehingga distribusi kelas pada data uji tetap tidak seimbang dan kesalahan pada kelas minoritas berdampak besar terhadap F1-*score*. Kedua, hasil t-SNE pada Gambar 2 menunjukkan tumpang tindih fitur antara kelas asma dan tidak asma, yang membatasi kemampuan model memisahkan kedua kelas. Ketiga, *threshold Information Gain* yang digunakan (0,0001) tergolong rendah sehingga masih mempertahankan fitur berkontribusi kecil yang berpotensi menambah *noise*. Keempat, parameter XGBoost ditentukan melalui percobaan manual, bukan tuning sistematis seperti *grid search*, sehingga kombinasi parameter yang diperoleh belum tentu optimal. Faktor-faktor ini menunjukkan bahwa performa model yang belum optimal lebih disebabkan oleh karakteristik dataset dan keterbatasan metodologi, bukan kegagalan *pipeline* yang diterapkan.

4. KESIMPULAN

Berdasarkan hasil dari seluruh pengujian yang telah dilakukan terhadap klasifikasi penyakit asma menggunakan empat skenario pendekatan, dapat disimpulkan bahwa penerapan seleksi fitur *Information Gain* dan SMOTE terbukti mampu meningkatkan peforma model XGBoost. Terlihat pada pengujian menggunakan data uji yang diseimbangkan dengan *RandomUnderSampler*, skenario *Information Gain* + SMOTE + XGBoost dengan rasio pembagian data 90:10 menghasilkan performa terbaik dengan akurasi 75%, presisi 87,5%, *recall* 58,33%, dan F1-*score* 70%. Sementara itu, pada pengujian menggunakan data uji asli yang tidak diseimbangkan, skenario terbaik juga diperoleh oleh *Information Gain* + SMOTE + XGBoost pada rasio 90:10 dengan akurasi 87,5%, presisi 21,88%, *recall* 58,33%, dan F1-*score* 31,82%. Meskipun nilai akurasi relatif tinggi, penurunan nilai presisi dan F1-*score* menunjukkan bahwa ketidakseimbangan kelas pada data uji masih memberikan dampak terhadap kemampuan model dalam mengidentifikasi kasus asma secara akurat.

Penelitian ini masih memiliki tantangan yang dapat dikembangkan lebih lanjut, yaitu performa model yang dipengaruhi oleh rasio pembagian data, serta penerapan SMOTE pada kondisi data yang tumpang tindih antar kelas sebagaimana ditunjukkan oleh visualisasi t-SNE, sehingga sebagian data sintetis yang dihasilkan kurang merepresentasikan karakteristik kelas asma secara jelas. Penelitian selanjutnya disarankan menerapkan teknik validasi silang seperti K-*Fold Cross Validation* agar evaluasi tidak bergantung pada satu pembagian data tertentu, serta mengeksplorasi varian SMOTE seperti *Borderline*-SMOTE atau ADASYN yang secara teori lebih terarah dalam menangani sampel pada wilayah perbatasan antar kelas dibandingkan SMOTE konvensional [28].

REFERENSI

[1] N. D. Rahmawati, I. L. Hilmi, and Salman, "Review of the Analysis of the Effectiveness and Risk of Aminophylline Toxicity in the Treatment of Asthma," *J. Pharm. Sci.*, vol. 6, no. 1, pp. 95–99, 2023, doi: https://doi.org/10.36490/journal-jps.com.v6i1.14.

[2] Z. J. Lee, M. R. Yang, and B. J. Hwang, "A Sustainable Approach to Asthma Diagnosis: Classification with Data Augmentation, Feature Selection, and Boosting Algorithm," *Diagnostics*, vol. 14, no. 7, 2024, doi: 10.3390/diagnostics14070723.

[3] B. S. Pansare, A. D. Kulkarni, and P. P. Pawar, "Data-Driven Approach for Asthma Classification : Ensemble Learning with Random Forest and XGBoost †," *Comput. Sci. Math. Forum*, vol. 12, no. 3, pp. 1–10, 2025, doi: https://doi.org/10.3390/cmsf2025012003.

[4] World Health Organization, "Asthma." Accessed: Apr. 18, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/asthma

[5] Sarman and S. Bahri, "Analisis Perbandingan Algoritma Naïve Bayes dengan SVM (Support Vektor Machine) Dalam Mendiagnosis Penyakit Asma Asthma," *J. Janitra Inform. dan Sist. Inf.*, vol. 5, no. 2, pp. 189–198, 2025, doi: 10.59395/m9qa0357.

[6] P. Ardhiyansyah *et al.*, "Klasifikasi Penyakit Asma Menggunakan Algoritma Decision Tree Pada Rapidminer," *J. Ilmu Komput. Dan Inform.*, vol. 2, no. 3, pp. 103–107, 2026, doi: https://jurnal.globalscients.com/index.php/jiki.1014 E-ISSN.

[7] A. U. Dullah, P. Utami, and Jumanto, "The Asthma Classification Using an Adaptive Boosting Model with SVM-SMOTE Sampling," *J. Inf. Syst. Explor. Res.*, vol. 3, no. 1, pp. 1–10, 2025, doi: 10.52465/joiser.v3i1.486.

[8] J. Sodiq and S. D. Permai, "Binary Classification of Asthma for the CAPS Pediatric Dataset in Malawi Using Machine Learning," *J. EMACS (Engineering, Math. Comput. Sci.*, vol. 7, no. 3, pp. 337–342, 2025, doi: 10.21512/emacsjournal.v6.

[9] A. F. B. Sajiwo, B. Rahmat, and A. Junaidi, "Klasifikasi Indeks Standar Pencemaran Udara (ISPU) Menggunakan Algoritma XGBoost Dengan Teknik Imbalanced Data (SMOTE)," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 12, no. 3, pp. 2190–2200, 2024, doi: http://dx.doi.org/10.23960/jitet.v12i3.4699.

[10] A. R. Illawati, A. I. Hadiana, and Melina, "Klasifikasi Penyakit Monkeypox dengan XGBoost dan SMOTE untuk Penanganan Data Tidak Seimbang," *J. Algoritm.*, vol. 22, no. 2, pp. 35–44, 2025, doi: 10.33364/algoritma/v.22-2.2349.

[11] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Inf.*, vol. 13, no. 10, 2022, doi: 10.3390/info13100475.

[12] Opitasari, F. Natsir, and E. S. Marsiani, "Klasifikasi Diagnosis untuk Penyakit Kanker Serviks Menggunakan Algoritma Extreme Gradient Boosting (XGBoost)," *J. Ilm. FIFO*, vol. 16, no. 1, p. 55, 2024, doi: 10.22441/fifo.2024.v16i1.006.

[13] R. Zizilia *et al.*, "Klasifikasi Penyakit Kanker Paru-Paru dengan Algoritma Extreme Gradient Boosting (XGBoost)dan Mutual Informationsebagai Metode Feature Selection," *Sist. J. Sist. Inf.*, vol. 14, no. 5, pp. 2540–9719, 2025, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[14] N. R. Muntiari, K. H. Hanif, Muliyadi, and Mufida, "Penanganan Ketidakseimbangan Data Pada Klasifikasi Penyakit Campak Menggunakan Kombinasi SMOTE Dan XGBoost," *JIKSTRA*, vol. 8, no. 01, pp. 42–51, 2026.

[15] M. A. Nugraha, M. I. Mazdadi, A. Farmadi, Muliadi, and T. H. Saragih, "Penyeimbang Kelas SMOTE Dan Seleksi Fitur Ensemble Filter Pada Support Vector Machine Untuk Klasifikasi Penyakit Liver," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1273–1284, 2023, doi: 10.25126/jtiik.2023107234.

[16] W. M. Putri, E. Budianita, F. Syafria, I. Afrianty, and K. Kunci, "Klasifikasi Penyakit Ginjal Kronis Menggunakan Information Gain Dan LVQ," *J. Inf. Syst. Manag.*, vol. 7, no. 1, pp. 48–56, 2025, doi: https://doi.org/10.24076/joism.2025v7i1.2102.

[17] G. F. Sijabat and W. A. E. Prabowo, "Studi Komparatif Model Machine Learning untuk Klasifikasi Penyakit Jantung dengan SMOTE pada Data Imbalanced," *JURIKOM (Jurnal Ris. Komputer)*, vol. 13, no. 1, pp. 398–409, 2026, doi: 10.30865/jurikom.v13i1.9485.

[18] N. C. Ramadhan, H. H. H, T. Rohana, and A. M. Siregar, "Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur Xgboost Untuk Klasifikasi Kanker Payudara," *TIN Terap. Inform. Nusant.*, vol. 5, no. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408.

[19] J. Pardede and R. Dwianto, "The Effect of Feature Selection on Machine Learning Classification," *Int. J. Informatics Vis.*, vol. 9, no. 4, pp. 1419–1429, 2025, doi: 10.62527/joiv.9.4.2926.

[20] N. Ulinnuha and A. Fanani, "Klasifikasi Status Drop Out Mahasiswa Menggunakan Naïve Bayes dengan Seleksi Fitur Information Gain," *Techno.Com*, vol. 22, no. 4, pp. 1014–1025, 2023, doi: 10.33633/tc.v22i4.9004.

[21] Devian, P. Nurul Sabrina, and A. Komarudin, "Prediksi Penyakit Diabetes Dengan Metode K-Nearest Neighbor (KNN) Dan Seleksi Fitur Information Gain," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 11320–11326, 2024, doi: 10.36040/jati.v8i6.11364.

[22] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[23] A. Yaqin and G. Ramadhani, "Penilaian Kredit Menggunakan Algoritma XGBoost dan Logistic Regression," *J. Inform. J. Pengemb. IT*, vol. 8, no. 1, pp. 4–10, 2022, doi: 10.30591/jpit.v8i1.4337.

[24] S. Murni, D. Widiyanto, and C. N. P. Dewi, "Klasifikasi Citra Penyakit Daun Kopi Arabika Menggunakan Support Vector Machine (SVM) Dengan Seleksi Fitur Information Gain," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, vol. 3, pp. 700–709, 2022.

[25] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "PENERAPAN SMOTE UNTUK MENGATASI IMBALANCE CLASS DALAM KLASIFIKASI KEPRIBADIAN MBTI MENGGUNAKAN NAIVE BAYES CLASSIFIER," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1033–1042, 2024, doi: 10.25126/jtiik.2024117989.

[26] D. Kurnia, M. Itqan Mazdadi, D. Kartini, R. Adi Nugroho, and F. Abadi, "Seleksi Fitur dengan Particle Swarm Optimization pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 5, pp. 1083–1094, 2023, doi: 10.25126/jtiik.2023107252.

[27] H. H. Sinaga and S. Agustian, "Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter," *J. Nas. Teknol. dan Sist. Inf.*, vol. 08, no. 03, pp. 107–114, 2022, doi: https://doi.org/10.25077/TEKNOSI.v8i3.2022.107-114.

[28] R. S. Abdulsadig and E. Rodriguez-villegas, "A comparative study in class imbalance mitigation when working with physiological signals," *Front. Digit. Heal.*, no. March, pp. 1–11, 2024, doi: 10.3389/fdgth.2024.1377165.