Journal of Artificial Intelligence and Software Engineering
Vol. 6, No. 2, Juni 2026, pp. 105~117
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v6i2.9354

Implementation of SMOTE and Information Gain Feature Selection in Learning Vector Quantization for Asthma Disease Classification

Diah Ayu Kinanti[1], Fitri Insani[2*], Novi Yanti[3], Muhammad Affandes[4]
[1,2,3,4] Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia

Informasi Artikel

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

Kata Kunci:

Learning Vector Quantization
SMOTE
Information Gain
Klasifikasi Penyakit Asma
Ketidakseimbangan Data

ABSTRAK

Asma adalah penyakit pernapasan akibat peradangan saluran udara di paru-paru yang menyebabkan penyempitan dan kesulitan bernapas. Prevalensinya terus meningkat secara global, sehingga diperlukan metode deteksi dini yang akurat. Masalah yang ditemukan dalam proses pengklasifikasian penyakit asma adalah distribusi kelas yang tidak seimbang pada dataset. Penelitian ini menerapkan algoritma *Learning Vector Quantization* (LVQ) yang dioptimalkan dengan seleksi fitur *Information Gain* dan teknik penyeimbangan data SMOTE untuk klasifikasi penyakit asma. Dataset penelitian mencakup 2.392 data pasien dengan 28 fitur dan 1 kelas target yang diperoleh dari *platform Kaggle*. Pengujian dilakukan pada lima skenario dengan tiga fungsi jarak *Euclidean, Chebyshev, Manhattan*, *learning rate* 0.001–0.005, dan rasio pembagian data 90:10, 80:20, serta 70:30. Hasil terbaik diperoleh pada skenario SMOTE, *Information Gain*, dan LVQ menggunakan fungsi jarak *Euclidean* dengan *learning rate* 0.004 dan rasio 90:10, menghasilkan akurasi 77.97%, *precision* 73.61%, *recall* 87.22% dan *F1-score* 79.84%. Penerapan SMOTE menjadi komponen penting karena tanpa SMOTE model gagal mengenali kelas asma, terbukti pada percobaan tanpa menggunakan SMOTE menghasilkan *precision, recall*, dan *F1-score* bernilai 0% meskipun akurasi mencapai 94–95%.

ABSTRACT

Asthma is a respiratory disease caused by inflammation of the airways in the lungs, leading to airway narrowing and breathing difficulties. Its prevalence continues to increase globally, highlighting the need for accurate early detection methods. One of the challenges in asthma classification is the imbalanced class distribution within the dataset. This study applies the Learning Vector Quantization (LVQ) algorithm optimized with Information Gain feature selection and the Synthetic Minority Over-sampling Technique (SMOTE) for asthma disease classification. The dataset consists of 2,392 patient records with 28 features and one target class, obtained from Kaggle. Experiments were conducted under five scenarios using three distance functions Euclidean, Chebyshev, and Manhattan with learning rates ranging from 0.001 to 0.005 and data-splitting ratios of 90:10, 80:20, and 70:30. The best performance was achieved in the scenario combining SMOTE, Information Gain, and LVQ using the Euclidean distance function, a learning rate of 0.004, and a 90:10 train-test split ratio. This configuration produced an accuracy of 77.97%, a precision of 73.61%, a recall of 87.22%, and an F1-score of 79.84%. The application of SMOTE proved to be a crucial component, as the model failed to recognize the asthma class without it. This was demonstrated by experiments conducted without SMOTE, which yielded precision, recall, and F1-score values of 0%, despite achieving an accuracy of 94–95%.

This is an open-access article under the CC BY-SA license

Penulis Koresponden

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

Cara sitasi IEEE:

D.A Kinanti, F. Insani, N. Yanti, dan M. Affandes, "Implementation of SMOTE and *Information Gain* Feature Selection in Learning Vector Quantization for Asthma Disease Classification," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 6, no. 2, p. 105-117, Juni 2026. doi:10.30811/jaise.v6i2.9354

1. PENDAHULUAN

Asma termasuk penyakit pernapasan dimana kondisi saluran udara di dalam paru-paru mengalami peradangan, sehingga menjadi lebih peka terhadap faktor-faktor pemicunya. Faktor-faktor ini menyebabkan saluran udara menyempit, mengurangi aliran udara, dan mengakibatkan kesulitan bernapas serta terdengar suara napas yang berbunyi seperti mengikik [1]. Berdasarkan laporan *World Health Organization* (WHO), pada tahun 2019 jumlah penderita asma di dunia diperkirakan mencapai 262 juta orang. Pada tahun yang sama, penyakit ini juga berkontribusi terhadap sekitar 455 ribu kasus kematian secara global [2]. Asma termasuk dalam sepuluh penyebab utama kematian dan penderitaan di Indonesia [3]. Asma memiliki tingkat prevalensi yang cukup beragam, mulai dari sekitar 1% di sejumlah wilayah Afrika dan Asia hingga melampaui 20% di sejumlah negara maju, seperti Australia, Selandia Baru, Inggris, dan Amerika Serikat [4]. Di Indonesia, angka prevalensi asma tercatat sebesar 4.5% dari seluruh jumlah penduduk, yang setara dengan lebih dari 12 juta jiwa [5].

Seiring berkembangnya teknologi secara pesat, mendorong pemanfaatan *machine learning* di berbagai bidang termasuk kesehatan. Metode ini memungkinkan analisis data dilakukan secara otomatis, sehingga mempermudah diagnosis penyakit dan proses pengambilan keputusan [6]. Penelitian terkait penerapan *machine learning* dalam klasifikasi penyakit asma telah dilakukan sebelumnya dengan menerapkan beberapa algoritma. Salah satu penelitian menggunakan *Adaptive Boosting* dengan *SVM-SMOTE* sebagai metode penyeimbang data menghasilkan akurasi mencapai 98.60%, dengan *precision, recall,* dan *F1-score* di atas 0.97 [7]. Studi lain membandingkan beberapa algoritma dengan menerapkan SMOTE sebagai metode penyeimbang data, pada algoritma *Naïve Bayes* memperoleh akurasi 96.99%, algoritma *Random Forest* memperoleh akurasi 97.35%, algoritma KNN memperoleh akurasi 97.50%, dan terakhir algoritma *Decision Tree* memperoleh akurasi 97.65% [8].

Salah satu pendekatan berbasis jaringan syaraf tiruan yang digunakan untuk klasifikasi data ke dalam kelas tertentu adalah *Learning Vector Quantization* (LVQ) [9]. Pada penelitian yang dilakukan Aziz et al. [10] menggunaan metode LVQ untuk mengklasifikasikan keluarga beresiko stunting. Penelitian tersebut menggunakan 7 neuron input, fungsi jarak *Chebyshev*, dengan *learning rate* 0.1, jumlah *epoch* 7, dan 30% data latih memberikan hasil yang optimal dengan tingkat akurasi mencapai 99.83%. Penelitian lainnya yang dilakukan Alamri et al. [11] membandingkan dua metode yaitu LVQ dan *Backpropagation* menunjukkan hasil bahwa dengan menggunakan metode LVQ memperoleh nilai akurasi sebesar 95.12%. Sedangkan menggunakan metode *Backpropagation* diperoleh akurasi sebesar 80.49%. Dari hasil perbandingan yang dilakukan, dapat diketahui bahwa metode LVQ lebih efektif dalam menyelesaikan permasalahan klasifikasi tingkat gizi balita di Kecamatan Sangkub dibandingkan metode *Backpropagation*.

Permasalahan data tidak seimbang, yaitu dominasi kelas tertentu terhadap kelas lainnya sering terjadi dalam data medis [12]. Data yang tidak seimbang dapat memengaruhi kinerja algoritma klasifikasi secara signifikan. Untuk mengatasi permasalahan tersebut, berbagai metode telah dikembangkan salah satunya adalah *Synthetic Minority Over-sampling Technique* (SMOTE). [13]. Dalam menangani ketidakseimbangan data, SMOTE bekerja dengan menghasilkan data sintetis baru yang berasal dari kelas minoritas. Teknik SMOTE yang diterapkan pada proses klasifikasi mampu mengurangi masalah *overfitting* [14]. Penelitian Muhammad Ibnu Choldun Rahmatullah et al. [13] mengenai klasifikasi data penyakit *stroke* dengan metode *Random Forest*, hasil pengujian menunjukkan bahwa sebelum diterapkan SMOTE model memperoleh akurasi sebesar 93.93%, tetapi gagal mengenali data pada kelas stroke yang ditandai dengan nilai *precision, recall, dan F1-score* sebesar 0%. Setelah diterapkan SMOTE, berhasil meningkatkan nilai *recall* sebesar 14.52%, *precision* sebesar 15.52%, dan *F1-score* sebesar 15.00%. Akan tetapi, peningkatan performa pada kelas minoritas tersebut menyebabkan akurasi keseluruhan model menurun menjadi 90.02%. Penelitian lain oleh Syukron et al. [14] pada kasus prediksi gagal jantung menunjukkan hasil penerapan metode SMOTE mampu meningkatkan performa akurasi pada enam algoritma klasifikasi. Pada algoritma C4.5, penerapan SMOTE meningkatkan akurasi sebesar 0.014. Algoritma *Naïve Bayes* menunjukkan peningkatan akurasi sebesar 0.004,

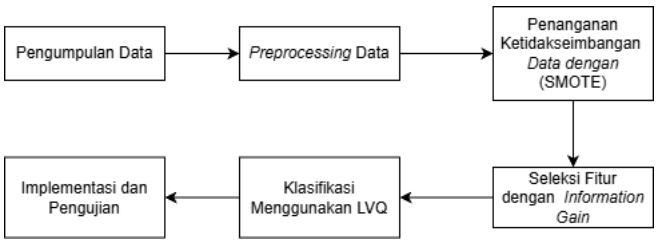
sedangkan pada algoritma *Neural Network* peningkatan yang diperoleh sebesar 0.010. Pada algoritma KNN dan *Bagging* masing-masing mengalami peningkatan akurasi sebesar 0.030. Sementara itu, Algoritma *Random Forest* menunjukkan peningkatan akurasi paling tinggi, yaitu sebesar 0.032.

Performa algoritma klasifikasi dapat ditingkatkan kinerjanya dengan salah satu cara, yaitu melalui penerapan seleksi fitur [15]. Seleksi fitur merupakan tahap pemilihan fitur yang tepat dari banyak fitur yang tersedia. Algoritma dapat memproses data lebih cepat jika fitur yang tidak relevan dihilangkan [16]. Terdapat beberapa penelitian terkait penerapan seleksi fitur *Information Gain*, yaitu penelitian Nada Tsawaabul Khair et al. [17] menerapkan seleksi fitur *Information Gain* dan algoritma *Backpropagation* pada klasifikasi jenis kelamin tulang tengkorak. Nilai ambang batas (*threshold*) yang digunakan pada tahap Seleksi fitur *Information Gain*, yaitu 0.01, 0.05, dan 0.1. Model kemudian diuji menggunakan kombinasi *learning rate* sebesar 0.1, 0.01, dan 0.001 untuk masing-masing *threshold*. Akurasi terbaik yang dihasilkan dalam penelitian ini mencapai 93.91%, yang diperoleh melalui metode *Information Gain* dengan *threshold* 0.01, arsitektur jaringan [79:119:1], *learning rate* 0.01, serta K = 20. Selanjutnya, penelitian oleh Anisa Fitri et al. [18] mengenai implementasi seleksi fitur *Information Gain* menggunakan *Support Vector Machine* untuk klasifikasi penyakit *stroke* berhasil memperoleh akurasi terbaik yaitu mencapai 90.51% pada kernel RBF dengan parameter *Cost* = 100 dan Gamma = 5 pada *threshold Information Gain* sebesar 0.0005.

Berdasarkan literatur tersebut, penerapan metode penyeimbangan data maupun seleksi fitur mampu meningkatkan kinerja algoritma klasifikasi. Namun, penelitian yang mengombinasikan SMOTE, *Information Gain*, dan LVQ untuk klasifikasi penyakit asma pada data yang tidak seimbang masih terbatas. Berbeda dengan penelitian sebelumnya yang umumnya hanya berfokus pada penerapan metode penyeimbangan data atau seleksi fitur pada algoritma tertentu, penelitian ini menerapkan kombinasi SMOTE, *Information Gain*, dan LVQ pada klasifikasi penyakit asma. Kontribusi penelitian ini adalah menganalisis pengaruh kombinasi SMOTE dan *Information Gain* terhadap kinerja algoritma LVQ pada klasifikasi penyakit asma dengan data yang tidak seimbang. Penelitian ini bertujuan untuk menerapkan dan menguji algoritma LVQ yang dioptimalkan menggunakan SMOTE dan *Information Gain* melalui beberapa skenario pengujian. Performa model dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*.

2. METODE

Pembangunan model LVQ untuk mendeteksi dini penyakit asma berbasis *machine learning* ini dilakukan melalui serangkaian tahapan penelitian yang saling terhubung dan berkelanjutan. Berikut skema dari tahapan penelitian yang dilakukan pada Gambar 1.



Gambar 1. Alur Penelitian Data Penyakit Asma

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder, berupa dataset tahun 2024 yang bersumber dari *Kaggle*. Data dapat diakses pada https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset. Dataset tersebut mencakup informasi kesehatan dari 2.392 pasien, dengan total 28 fitur dan 1 kelas target. Distribusi kelas diagnosis pada dataset penyakit asma ini memiliki ketidakseimbangan dengan 2.268 data mewakili kelas 0 yang merepresentasikan pasien tidak mengalami penyakit asma dan 124 data mewakili kelas 1 yang merepresentasikan pasien mengalami asma. Beberapa fitur yang terdapat pada dataset dapat dilihat pada Tabel 1.

Tabel 1. Fitur yang Terdapat Pada Dataset Asma

No	Fitur	Deskripsi
1.	*PatientID*	ID Pasien
2.	*Age*	Usia pasien
3.	*Gender*	Jenis kelamin pasien
4.	*Ethnicity*	Etnis pasien
5.	*EducationLevel*	Tingkat pendidikan pasien
…….	……	…….
29	*DoctorInCharge*	Dokter yang bertanggungjawab

2.2 *Preprocessing* Data

2.2.1 Seleksi Data

Tahap seleksi data adalah tahap memilih atau menyeleksi sekumpulan data yang dilakukan sebelum memasuki tahap penggalian informasi lebih dalam [19] [20]. Seleksi data merupakan tahapan yang dilakukan untuk memilih data dan fitur yang memiliki keterkaitan dengan tujuan penelitian dari seluruh data yang tersedia, sehingga efektivitas pengolahan data dan kualitas hasil yang diperoleh dapat ditingkatkan.

2.2.2 Pembersihan Data

Proses ini dilakukan untuk memeriksa dan membersihkan data dari masalah seperti data duplikat dan data yang hilang [17]. Langkah ini bertujuan untuk mencegah bias akibat data yang hilang sekaligus memastikan semua fitur siap digunakan dalam pelatihan model.

2.2.3 Transformasi Data

Transformasi adalah proses mengubah bentuk, skala, atau representasi data ke format yang lebih sesuai agar dapat diolah dan dianalisis dengan baik oleh algoritma *machine learning*. Salah satu proses transformasi data yang dilakukan adalah *encoding* fitur kategorikal menggunakan *One Hot Encoding*. Pada tahap berikutnya, dilakukan normalisasi data menggunakan metode *MinMax Normalization* untuk menyamakan skala nilai setiap fitur ke rentang 0–1. Normalisasi *MinMax* dihitung dengan menggunakan rumus yang ditunjukkan pada Persamaan 1.

.

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

Dimana (x') merupakan hasil normalisasi, (x_i) adalah data yang akan dinormalisasi, (min(x)) menunjukkan nilai terkecil pada suatu fitur, sedangkan (\max(x)) merupakan nilai terbesar pada fitur tersebut.

2.3 *Synthetic Minority Over-sampling Technique* (SMOTE)

SMOTE merupakan metode *oversampling* yang diterapkan untuk menangani permasalahan ketidakseimbangan kelas dalam dataset [21]. Cara menghasilkan sampel sintetis yaitu dari kelas yang memiliki jumlah data paling sedikit. Jumlah sampel data akan ditingkatkan hingga seimbang dengan kelas yang memiliki jumlah data terbanyak, sehingga distribusi data menjadi lebih seimbang [7]. Berikut merupakan rumus yang digunakan dalam metode SMOTE :

$$\left(\boldsymbol{X_{syn}}\right) = \boldsymbol{X_i} + (\boldsymbol{X_{knn}} - \boldsymbol{X_i}) \times \boldsymbol{\delta} \quad (2)$$

Pada metode SMOTE, proses diawali dengan memilih satu sampel dari kelas minoritas yang dinotasikan sebagai X_i. Selanjutnya, dicari k tetangga terdekat (*K-Nearest Neighbors*) yang berasal dari kelas yang sama. Salah satu tetangga terdekat yang diperoleh dipilih sebagai X_{knn}. Kemudian, dihitung selisih antara $X_i$dan X_{knn}, yang selanjutnya dikalikan dengan nilai acak δ dalam rentang 0 hingga 1. Hasil perhitungan tersebut ditambahkan ke X_i untuk menghasilkan sampel sintetis baru $\left(X_{syn}\right)$ yang digunakan sebagai tambahan data pada kelas minoritas.

2.4 Seleksi Fitur *Information Gain*

Berbagai bidang penelitian, termasuk klasifikasi teks, analisis genomik, deteksi intrusi, dan bioinformatika, telah banyak menggunakan teknik seleksi fitur. *Informasi Gain* adalah teknik seleksi fitur yang umum digunakan [17]. *Information Gain* digunakan dalam proses seleksi fitur untuk memilih fitur paling relevan sehingga dapat mengurangi dimensi data dan meningkatkan hasil prediksi [22]. Berikut adalah langkah-langkah perhitungan *Information Gain* :

Perhitungan nilai *entropy* dilakukan untuk mengetahui tingkat ketidakpastian suatu kelas, yang diperoleh berdasarkan probabilitas kemunculan setiap fitur atau kategori dalam dataset.

$$Entropy(S) = \sum_{i=1}^{n} - p_i \log_2 p_i \quad (3)$$

Menghitung *Information Gain* menerapkan Persamaan (4).

$$Gain(S,A) = Entropy(S) - \sum_{values\ (A)} \frac{|sv|}{|s|} Entropy\ (S_v) \quad (4)$$

Pada Persamaan (4), S menyatakan himpunan data yang digunakan dalam proses klasifikasi. Simbol $|\ S\ |$ menunjukkan jumlah seluruh sampel pada dataset, sedangkan p_i menyatakan proporsi atau probabilitas kemunculan kelas ke-i dalam data. Atribut yang akan dievaluasi dinotasikan dengan A, sementara $Values(A)$ merupakan himpunan seluruh nilai yang dimiliki oleh atribut tersebut. Selanjutnya, S_v menyatakan subset data dengan nilai atribut Asama dengan v, dan $|\ S_v\ |$ menunjukkan jumlah sampel pada subset tersebut. Adapun $Entropy(S_v)$ merupakan nilai entropy dari subset data S_v.

2.5 *Learning Vector Quantization* (LVQ)

LVQ adalah metode pelatihan yang digunakan untuk proses pembelajaran terawasi (*supervised learning*) pada lapisan kompetitif dengan arsitektur jaringan berlayer tunggal (*single layer*) [23]. Lapisan kompetitif secara otomatis mempelajari karakteristik vektor input selama proses pelatihan sehingga dapat melakukan klasifikasi berdasarkan pola yang telah dipelajari. Kelas yang dihasilkan ditentukan berdasarkan jarak antar vektor input. Langkah-langkah perhitungan pada algoritma LVQ dijelaskan sebagai berikut:

a. Pada tahap awal menentukan parameter yaitu, nilai *learning rate*(α), jumlah iterasi (*maxEpoch*), pengurang *learning rate* (*dec α*), minimum α dan menentukan vektor bobot awal kelas (w).
b. Masukkan data *input* serta kelas atau kategori target.
c. Kerjakan apabila ($epoch \leq max\ epoch\ dan\ \alpha \geq min\ \alpha$):
    1. *Epoch* = *epoch* + 1;
    2. Hitung jarak minimum, dapat menggunakan perhitungan jarak berikut
       Rumus jarak *Euclidean* antara vektor input X dengan vektor bobot W ke-i
       $$\mathbf{d(X, W_i)} = \sqrt{\sum_{\mathbf{j=1}}^{\mathbf{j}} \mathbf{(x_j - w_{ij})^2}} \quad (5)$$
       Rumus jarak *Manhattan*
       $$\mathbf{d(X, W_i)} = \sum_{\mathbf{j=1}} \mathbf{|\ X_j - W_{ij}\ |^2} \quad (6)$$
       Rumus jarak *Chebyshev*
       $$\mathbf{d(X, W_i)} = \mathbf{max}(\mathbf{|\ X_j - W_{ij}\ |})^2 \quad (7)$$
d. Memperbarui nilai bobot berdasarkan hasil perhitungan.:
   Jika T = Cj maka:
   W_j (baru) = W_j (lama) + a (X_i - W_j (lama)) (8)
   Jika T ≠ Cj maka:
   W_j (baru) = W_j (lama) - a (X_i - W_j (lama)) (9)
e. Penurunan nilai α dilakukan menggunakan persamaan berikut:
   $$a\ (baru) = a(lama) - (a * \text{dec}\ \alpha) \quad (10)$$
f. Tahap pelatihan berhenti saat maksimum *epoch* tercapai atau *learning rate* (α) mencapai nilai minimum.
g. Setelah seluruh proses pelatihan dilakukan, diperoleh nilai bobot akhir (w).

2.6 Implementasi dan Pengujian

Pada penelitian ini, tahap implementasi dilakukan dengan mengombinasikan algoritma LVQ, metode seleksi fitur *Information Gain*, dan teknik SMOTE. Proses implementasi menggunakan bahasa pemrograman *Python* melalui platform *Google Colab*. Selanjutnya, model yang telah dilatih dievaluasi menggunakan data uji untuk mengukur tingkat kinerjanya. Salah satu metode yang dapat diterapkan untuk mengevaluasi kinerja model adalah *confusion matrix*, yang berfungsi untuk menilai kemampuan model dalam melakukan klasifikasi. Terdapat empat istilah nilai yang digunakan dalam pengukuran kinerja klasifikasi, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [24]. Berikut adalah beberapa langkah evaluasi yang dilakukan melalui confusion matrix :

a. Akurasi
   $$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$
b. Presisi
   $$Precision = \frac{TP}{TP + FP} \quad (12)$$

c. Recall

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

d. F1-score

$$F1 = 2.\frac{Precision\ .\ Recall}{Precision + Recall} \quad (14)$$

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan kumpulan data kesehatan yang terdiri dari 2.392 pasien dengan total 28 fitur dan 1 kelas target. Dataset tersebut terdiri atas dua kategori kelas, yaitu 0 merepresentasikan pasien yang tidak mengalami asma dan 1 merepresentasikan pasien yang mengalami asma. Dataset awal ditunjukkan pada Tabel 2.

Tabel 2. Dataset Awal Penyakit Asma

Patient ID	Age	Gender	Ethnicity	Education Level	BMI	Smoking	Physical Activity	…….	Diagnosis
5034	63	0	1	0	15.848744	0	0.894448	…….	0
5035	26	1	2	2	22.757042	0	5.897329	…….	0
5036	57	0	2	1	18.395396	0	6.739367	…….	0
…….	…….	…….	…….	…….	…….	…….	…….	…….	…….
7425	26	1	0	0	28.123021	1	1.613138	…….	0

3.2. *Preprocessing* Data

3.2.1. Seleksi Data

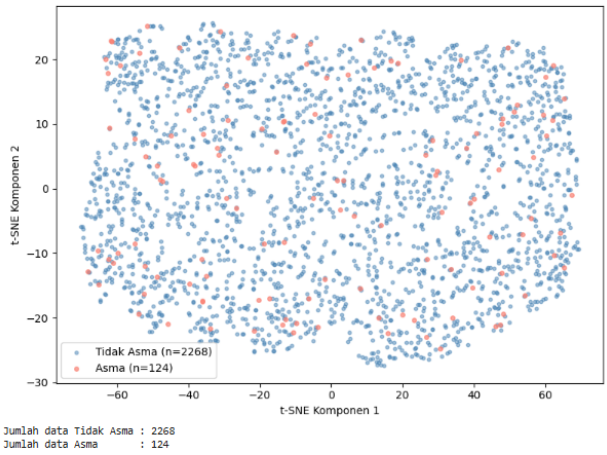
Pada proses ini, dilakukan penghapusan dua fitur yang tidak memiliki keterkaitan dengan proses klasifikasi yaitu fitur *PatientID* dan *DoctorInCharge* karena fitur tersebut hanya berfungsi sebagai nomor identitas data dan informasi dokter yang bertanggung jawab yang bersifat rahasia serta tidak berpengaruh terhadap hasil klasifikasi. Setelah dilakukan proses penghapusan, jumlah fitur berkurang dari 28 fitur dan 1 kelas target menjadi 26 fitur dan 1 kelas target.

3.2.2. Pembersihan Data

Proses ini dilakukan untuk memeriksa dataset dengan memastikan bahwa tidak terdapat *missing value*, data duplikat, maupun *outlier*. Berdasarkan hasil pemeriksaan terhadap dataset penyakit asma, tidak ditemukan adanya *missing values*, data duplikat, maupun *outlier* pada seluruh fitur yang tersedia.

3.2.3 Analisis Persebaran Data

Sebelum dilakukan transformasi data, dilakukan analisis persebaran data untuk mengetahui karakteristik distribusi kelas pada dataset. Visualisasi persebaran data dilakukan menggunakan metode t-SNE (*t-Distributed Stochastic Neighbor Embedding*) yang ditampilkan pada Gambar 2.



Gambar 2. Visualisasi Persebaran Data Menggunakan Metode t-SNE

Berdasarkan Gambar 2. visualisasi t-SNE menunjukkan adanya ketidakseimbangan kelas yang ekstrem antara kelas tidak asma dan kelas asma, serta terdapat *class overlap* dimana kedua kelas tersebar bercampur tanpa batas pemisah yang jelas akibat kemiripan karakteristik fitur. Kondisi ini menjadi dasar penerapan SMOTE untuk menyeimbangkan distribusi data sebelum proses klasifikasi.

3.2.4. Transformasi Data

Pada tahap ini dilakukan proses *encoding* pada fitur *Ethnicity* menggunakan *One Hot Encoding* selanjutnya dilakukan proses normalisasi pada fitur yang bernilai kontinu menggunakan *MinMax Normalization*. Fitur *Ethnicity* pada dataset direpresentasikan dalam bentuk angka yaitu 0 untuk *Caucasian*, 1 untuk *African American*, 2 untuk *Asian*, dan 3 untuk *Other*. Oleh karena itu, fitur ini bersifat kategorikal nominal yang tidak memiliki urutan matematis, penggunaan angka secara langsung dapat menyebabkan model salah menginterpretasikan bahwa satu kategori lebih besar dari kategori lainnya oleh karena itu dilakukan tahap *encoding*. Hasil *One Hot Encoding* dapat dilihat pada Tabel 3.

Tabel 3. Hasil *One Hot Encoding*

No	Ethnicity_Caucasian	Ethnicity_African American	Ethnicity_Asian	Ethnicity_Other
1.	0	1	0	0
2.	0	0	1	0
3.	0	0	1	0
.....	.....	.....	.....	.....
2.392	1	0	0	0

Setelah proses *One Hot Encoding* diterapkan, jumlah fitur bertambah dari 26 fitur menjadi 29 fitur. Menghasilkan empat fitur biner baru yaitu *Ethnicity_Caucasian, Ethnicity_AfricanAmerican*, *Ethnicity_Asian*, dan *Ethnicity_Other*. Keempat fitur baru tersebut merepresentasikan nilai kategorikal pada fitur *Ethnicity* dalam bentuk numerik biner, dimana nilai 1 menunjukkan bahwa pasien termasuk dalam kategori tersebut dan nilai 0 menunjukkan sebaliknya. Selanjutnya dilakukan normalisasi menggunakan *MinMax Normalization*, fitur yang dinormalisasi adalah *Age, EducationLevel, BMI, PhysicalActivity, DietQuality, SleepQuality, PollutionExposure, PollenExposure, Dust Exposure, Lung Function FEV, dan Lung Function FVC* yang dapat dilihat pada Tabel 4.

Tabel 4. Daftar Fitur Kontinu

No	Age	Education Level	BMI	Physical Activity	Diet Quality	Sleep Quality	Pollution Exposure	Pollen Exposure	.....	Lung Function FVC
1.	63	0	15.84874	8.94448	5.48869	8.70100	7.38848	2.85557	.....	4.94120
2.	26	2	22.75704	5.89732	6.34101	5.15396	1.96983	7.45766	.....	1.70239
3.	57	1	18.39539	6.73936	9.19623	6.84064	1.46059	1.44818	.....	5.02255
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
2.392	26	0	28.12302	1.61313	7.41287	8.51225	3.23170	3.87402	.....	2.45328

Tabel 4. menunjukkan fitur kontinu yang akan dinormalisasi, fitur-fitur tersebut memiliki skala nilai yang berbeda-beda sebagai contoh, fitur *Age* memiliki rentang nilai puluhan, sedangkan beberapa fitur lainnya memiliki rentang nilai yang lebih kecil. Perbedaan skala ini dapat menyebabkan fitur dengan nilai yang lebih besar mendominasi proses perhitungan jarak pada algoritma LVQ. Oleh karena itu, diperlukan tahap normalisasi data untuk menyamakan skala seluruh fitur sehingga setiap atribut dapat memberikan kontribusi yang seimbang dalam proses pembelajaran dan klasifikasi. Berikutnya hasil dari normalisasi ditunjukkan pada Tabel 5.

Tabel 5. Hasil *MinMax Normalization* Fitur Kontinu

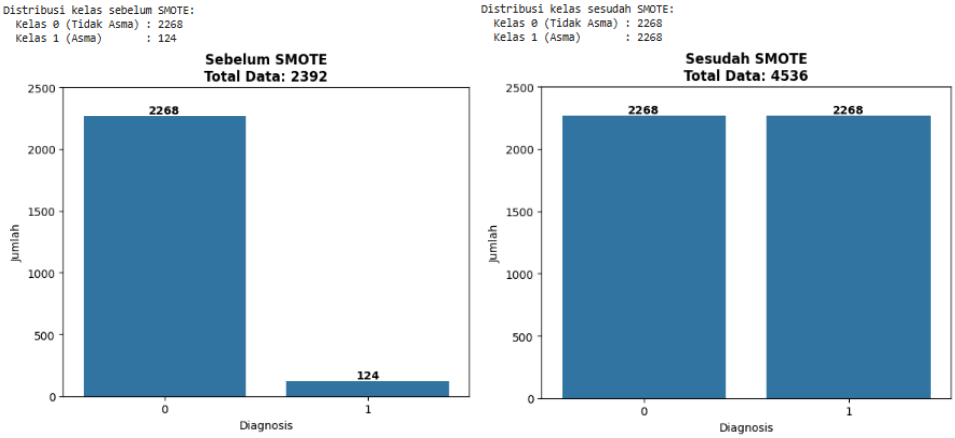
No	Age	Education Level	BMI	Physical Activity	Diet Quality	Sleep Quality	Pollution Exposure	Pollen Exposure	.....	Lung Function FVC
1.	0.7837	0	0.03273	0.08932	0.54873	0.78394	0.73889	0.28552	.....	0.76480
2.	0.2837	0.66666	0.30958	0.58990	0.63399	0.19225	0.19692	0.74578	.....	0.04497
3.	0.7027	0.33333	0.13479	0.67416	0.91960	0.47361	0.14598	0.14476	.....	0.78288
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
2.392	0.2837	0	0.52461	0.16123	0.74121	0.75245	0.32313	0.38737	.....	0.21186

Tabel 5. menampilkan hasil normalisasi fitur kontinu menggunakan metode *MinMax Normalization*. Metode ini mengubah nilai setiap fitur ke dalam rentang 0 hingga 1 tanpa menghilangkan karakteristik atau hubungan antar data. Berdasarkan hasil normalisasi, seluruh fitur telah berhasil ditransformasikan ke skala

yang sama sehingga seluruh fitur memiliki rentang nilai yang seragam dan tidak terdapat perbedaan skala yang signifikan antar fitur.

3.3. SMOTE

Pada tahapan ini dilakukan penyeimbangan kelas menggunakan SMOTE karena terdapat ketidakseimbangan data yang ekstrem antara kelas. Dimana kelas 0 (tidak asma) berjumlah 2.268 data dan kelas 1 (asma) berjumlah 124 data yang dapat menyebabkan model bias terhadap kelas mayoritas. SMOTE bekerja dengan membentuk data sintetis baru pada kelas minoritas berdasarkan sampel minoritas dan tetangga terdekatnya menggunakan bilangan acak sesuai persamaan (2). Visualisasi distribusi kelas sebelum dan sesudah penerapan SMOTE ditunjukkan pada Gambar 3.



Gambar 3. Distribusi Kelas Sebelum dan Sesudah Penyeimbangan Data

Berdasarkan Gambar 3. sebelum diterapkan SMOTE terdapat ketidakseimbangan kelas yang signifikan dimana kelas 0 (tidak asma) berjumlah 2.268 data sedangkan kelas 1 (asma) hanya berjumlah 124 data. Kondisi ini menunjukkan rasio ketidakseimbangan data yang ekstrem, sehingga berpotensi menyebabkan model bias terhadap kelas mayoritas. Untuk mengatasi hal tersebut, SMOTE diterapkan dengan cara membuat data sintetis baru pada kelas minoritas melalui interpolasi antar data yang berdekatan, sehingga jumlah kelas 1 (Asma) meningkat dari 124 menjadi 2.268 data. Hasilnya total data meningkat dari 2.392 menjadi 4.536 data dengan distribusi kelas yang seimbang.

3.4. Seleksi Fitur *Information Gain*

Tahap awal pada seleksi fitur *Information Gain* adalah menghitung nilai *entropy*, yang diperoleh dari Persamaan (3). Selanjutnya dilakukan perhitungan nilai *entropy* pada setiap fitur, lalu melakukan perhitungan nilai *Information Gain* dengan Persamaan (4) pada setiap fitur guna menentukan tingkat relevansi fitur terhadap target kelas. Dilakukan beberapa percobaan menggunakan *threshold* yang dapat dilihat pada tabel 6.

Tabel 6. *Threshold* yang digunakan pada *Information Gain*

Skenario	*Threshold* yang Digunakan
Information Gain dan LVQ	0.0007, 0.0001
SMOTE, *Information Gain*, dan LVQ	0.01, 0.007, 0.001
Information Gain, SMOTE, dan LVQ	0.0007, 0.0001

Berdasarkan beberapa kombinasi skenario percobaan, dengan beberapa *threshold* tersebut didapatkan parameter terbaik pada *threshold* 0.0001 untuk skenario *Information Gain* dan LVQ serta skenario *Information Gain*, SMOTE, dan LVQ. Kemudian pada skenario SMOTE, *Information Gain*, dan LVQ diperoleh *threshold* terbaik yaitu 0.007. Pada kedua *threshold* tersebut mendapatkan fitur terpilih sebanyak 22 fitur.

3.5. Implementasi dan Pengujian

Pengujian metode LVQ menggunakan rasio pembagian data latih dan data uji sebesar 90:10, 80:20, dan 70:30. Fungsi jarak yang digunakan yaitu *Euclidean*, *Chebyshev*, dan *Manhattan*. Parameter *learning rate* yang digunakan antara lain 0.001, 0.002, 0.003, 0.004, dan 0.005, dengan nilai minimum *learning rate* 0.0001 dan pengurangan alpha sebesar 0.001. Evaluasi performa model akan dihitung berdasarkan metode *Confusion Matrix* untuk memperoleh akurasi, *precision*, *recall,* dan *F1-score* menggunakan persamaan (11) hingga (14). Model terbaik dipilih berdasarkan nilai *F1-score* tertinggi, karena metrik tersebut mampu menggambarkan

performa model secara menyeluruh melalui keseimbangan antara *precision* dan *recall* . Semakin tinggi nilai *F1-score*, semakin baik kemampuan model dalam menangani klasifikasi pada data yang tidak seimbang. Berikut kombinasi skenario pengujian yang dilakukan:

a. Pengujian menggunakan LVQ
b. Pengujian menggunakan *Information Gain* dengan *threshold* 0.0001 dan LVQ
c. Pengujian menggunakan SMOTE dan LVQ
d. Pengujian menggunakan SMOTE, *Information Gain* dengan *threshold* 0.007 dan LVQ
e. Pengujian menggunakan *Information Gain* dengan *threshold* 0.0001, SMOTE, dan LVQ

Skenario pengujian pertama yaitu menggunakan algoritma LVQ tanpa melibatkan teknik penyeimbangan data ataupun seleksi fitur. Hasil pengujian disajikan pada Tabel 7.

Tabel 7. Hasil Pengujian Skenario LVQ

Skenario	Jarak	Rasio Split	*Learning rate*Terbaik	Akurasi	*Precision*	*Recall*	*F1-score*
LVQ	*Euclidean*	90:10	-	94.58%	0%	0%	0%
		80:20	-	94.78%	0%	0%	0%
		70:30	-	94.71%	0%	0%	0%
	Chebyshev	90:10	-	95.00%	0%	0%	0%
		80:20	-	94.78%	0%	0%	0%
		70:30	-	94.85%	0%	0%	0%
	Manhattan	90:10	-	94.58%	0%	0%	0%
		80:20	-	94.78%	0%	0%	0%
		70:30	-	94.71%	0%	0%	0%

Pengujian dengan skenario menggunakan algoritma LVQ tanpa teknik penyeimbangan data ataupun seleksi fitur, memperoleh akurasi berkisar antara 94.58%–95.00%. Namun hasil ini tidak dapat dijadikan ukuran keberhasilan model, karena nilai *precision, recall, dan F1-score* untuk kelas asma bernilai 0%. Artinya model sama sekali tidak mampu mendeteksi pasien yang menderita asma dan hanya memprediksi semua data sebagai kelas tidak asma (kelas mayoritas). Selanjutnya pengujian skenario kedua, menggunakan seleksi fitur *Information Gain* sebelum masuk tahap klasifikasi dengan LVQ, hasil pengujian dapat dilihat pada Tabel 8.

Tabel 8. Hasil Pengujian Skenario *Information Gain* dengan *threshold 0.0001* dan LVQ

Skenario	Jarak	Rasio Split	*Learning rate*Terbaik	Akurasi	*Precision*	*Recall*	*F1-score*
IG dan LVQ	*Euclidean*	90:10	-	94.17%	0%	0%	0%
		80:20	-	94.78%	0%	0%	0%
		70:30	-	94.71%	0%	0%	0%
	Chebyshev	90:10	-	95.00%	0%	0%	0%
		80:20	-	94.78%	0%	0%	0%
		70:30	-	94.85%	0%	0%	0%
	Manhattan	90:10	-	94.58%	0%	0%	0%
		80:20	-	94.78%	0%	0%	0%
		70:30	-	94.85%	0%	0%	0%

Hasil pengujian menunjukkan bahwa penambahan seleksi fitur *Information Gain* tidak meningkatkan performa model meskipun jumlah fitur telah diseleksi menjadi fitur-fitur yang lebih relevan berdasarkan nilai *Information Gain*, dengan akurasi berkisar 94.17%–95.00% namun *precision, recall,* dan *F1-score* tetap bernilai 0% pada seluruh kombinasi fungsi jarak dan rasio yang diuji. Hal ini menunjukkan bahwa permasalahan ketidakseimbangan kelas masih menyebabkan model hanya memprediksi kelas mayoritas tanpa mampu mengenali kelas asma sama sekali. Kemudian pengujian skenario ketiga, menggunakan SMOTE untuk menyeimbangkan data sebelum masuk ke tahap klasifikasi dengan LVQ, hasil pengujian skenario tersebut dapat dilihat pada Tabel 9.

Tabel 9. Hasil Pengujian Skenario SMOTE dan LVQ

Skenario	Jarak	Rasio Split	*Learning rate* Terbaik	Akurasi	*Precision*	*Recall*	*F1-score*
SMOTE + LVQ	*Euclidean*	90:10	0.001	76.43%	72.56%	85.02%	78.30%
		80:20	0.003	75.22%	71.40%	84.14%	77.25%
		70:30	0.005	74.06%	70.72%	82.06%	75.97%
	Chebyshev	90:10	0.005	70.48%	70.13%	71.37%	70.74%
		80:20	0.004	66.19%	63.74%	75.11%	68.96%
		70:30	0.002	67.52%	78.20%	48.53%	59.89%
	Manhattan	90:10	0.001	68.50%	61.73%	97.36%	75.56%
		80:20	0.001	65.97%	59.89%	96.70%	73.97%
		70:30	0.003	65.54%	59.65%	95.88%	73.55%

Hasil pengujian skenario SMOTE dan LVQ menunjukkan peningkatan performa yang signifikan dibandingkan skenario sebelumnya, dimana penerapan SMOTE terbukti efektif menangani ketidakseimbangan kelas sehingga model mampu mengenali kelas asma dengan baik. Hasil terbaik diperoleh pada fungsi jarak *Euclidean* dengan rasio 90:10 dan learning rate 0.001, menghasilkan *F1-score* 78.30%, *precision* 72.56%, dan *recall* 85.02%. Sedangkan pada pengujian menggunakan fungsi jarak *Chebyshev F1-score* tertinggi yang diperoleh sebesar 70.74% dengan *learning rate* 0.005 dan rasio pembagian data 90:10. Pada penggunaan fungsi jarak *Manhattan*, *F1-score* tertinggi yang diperoleh sebesar 75.56% dengan *learning rate* 0.001 serta rasio pembagian data 90:10. Berikutnya pengujian skenario keempat, menggunakan SMOTE, *Information Gain* dengan *threshold* 0.007 dan LVQ disajikan pada Tabel 10.

Tabel 10. Hasil Pengujian Skenario SMOTE, *Information Gain* dengan *threshold* 0.007 dan LVQ

Skenario	Jarak	Rasio Split	*Learning rate* Terbaik	Akurasi	*Precision*	*Recall*	*F1-score*
SMOTE, *Information Gain*, dan LVQ	*Euclidean*	90:10	0.004	77.97%	73.61%	87.22%	79.84%
		80:20	0.001	75.11%	71.51%	83.48%	77.03%
		70:30	0.001	73.84%	70.93%	80.74%	75.52%
	Chebyshev	90:10	0.003	68.28%	81.68%	47.14%	59.78%
		80:20	0.001	66.30%	81.09%	42.51%	55.78%
		70:30	0.001	65.47%	79.33%	41.76%	54.72%
	Manhattan	90:10	0.003	67.84%	61.54%	95.15%	74.74%
		80:20	0.004	60.06%	94.71%	73.50%	65.86%
		70:30	0.001	60.21%	91.47%	72.62%	65.54%

Pengujian menggunakan SMOTE, *Information Gain* dengan *threshold* 0.007, dan LVQ memperoleh *F1-score* tertinggi sebesar 79.84% dengan *precision* 73.61%, dan *recall* 87.22% pada penggunaan fungsi jarak *Euclidean* dengan *learning rate* 0.004 serta rasio pembagian data 90:10. Sementara itu, pada penggunaan fungsi jarak *Chebyshev*, *F1-score* tertinggi yang diperoleh sebesar 59.78% dengan *learning rate* 0.003 dan rasio pembagian data 90:10. Adapun pada penggunaan fungsi jarak *Manhattan*, *F1-score* tertinggi yang diperoleh sebesar 74.74% dengan *learning rate* 0.003 serta rasio pembagian data 90:10. Adapun pengujian skenario kelima, menggunakan *Information Gain* dengan *threshold* 0.0001, SMOTE dan LVQ dapat dilihat pada Tabel 11.

Tabel 11. Hasil Pengujian *Information Gain* dengan *threshold* 0.0001, SMOTE dan LVQ

Skenario	Jarak	Rasio Split	*Learning rate* Terbaik	Akurasi	*Precision*	*Recall*	*F1-score*
Information Gain, SMOTE, dan LVQ	*Euclidean*	90:10	0.001	70.04%	68.42%	74.45%	71.31%
		80:20	0.005	66.96%	64.42%	75.77%	69.64%
		70:30	0.005	66.64%	65.31%	70.88%	67.98%
	Chebyshev	90:10	0.005	65.20%	74.82%	45.81%	56.83%
		80:20	0.001	59.03%	72.78%	28.85%	41.32%
		70:30	0.001	59.07%	74.50%	27.50%	40.17%
	Manhattan	90:10	0.001	64.76%	59.88%	89.43%	71.73%
		80:20	0.001	61.89%	57.94%	86.78%	69.49%
		70:30	0.005	62.45%	58.37%	86.62%	69.75%

Pengujian menggunakan *Information Gain* dengan *threshold* 0.0001, SMOTE, dan LVQ memperoleh *F1-score* tertinggi sebesar 71.31% pada fungsi jarak *Euclidean* dengan learning rate 0.001 dan rasio pembagian data 90:10, diikuti fungsi jarak *Manhattan* sebesar 71.73% dengan learning rate 0.001 dan rasio pembagian data 90:10, serta fungsi jarak *Chebyshev* sebesar 56.83% dengan learning rate 0.005 dan rasio pembagian data 90:10. Perbandingan performa berdasarkan nilai *F1-score* tertinggi ditampilkan pada Tabel 12.

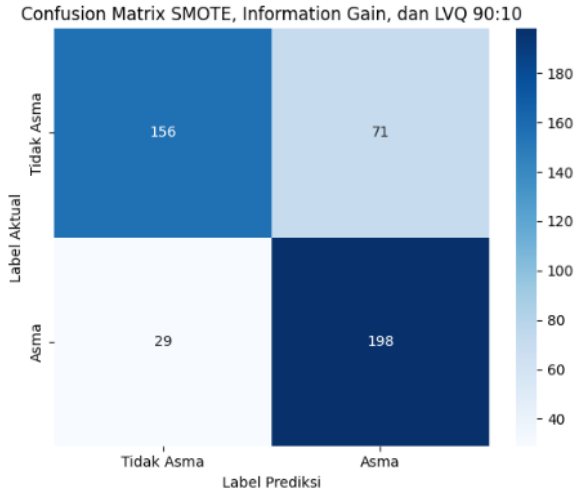
Tabel 12. Perbandingan Performa Berdasarkan Nilai *F1-score* Tertinggi

Skenario	Jarak	Rasio Split	*Learning rate* Terbaik	Akurasi	*Precision*	*Recall*	*F1-score*
LVQ	*Chebyshev*	90:10	-	95.00%	0%	0%	0%
Information Gain dan LVQ	*Chebyshev*	90:10	-	95.00%	0%	0%	0%
SMOTE dan LVQ	*Euclidean*	90:10	0.001	76.43%	72.56%	85.02%	78.30%
SMOTE, *Information Gain*, dan LVQ	*Euclidean*	90:10	0.004	77.97%	73.61%	87.22%	79.84%

Information Gain, SMOTE, dan LVQ	*Manhattan*	90:10	0.001	64.76%	59.88%	89.43%	71.73%

Berdasarkan keseluruhan hasil pengujian yang ditampilkan pada Tabel 12, pengujian dengan skenario keempat yaitu model LVQ dengan penerapan SMOTE terlebih dahulu, kemudian dilanjutkan dengan seleksi fitur *Information Gain*, menghasilkan nilai *F1-score* tertinggi yaitu 79.84%. Selain itu penggunaan fungsi jarak *Euclidean* terbukti menghasilkan performa terbaik dibandingkan fungsi jarak *Manhattan* dan *Chebyshev*. Hal ini menunjukkan bahwa kombinasi SMOTE sebagai penanganan ketidakseimbangan kelas dan *Information Gain* sebagai seleksi fitur terbukti mampu meningkatkan kemampuan model LVQ dalam mengklasifikasikan penyakit asma dibandingkan dengan pengujian yang tidak menggunakan kedua metode tersebut. Skenario yang menerapkan SMOTE sebelum *Information Gain* menghasilkan performa yang lebih baik dibandingkan skenario yang menerapkan *Information Gain* sebelum SMOTE. Hal ini disebabkan ketika SMOTE diterapkan terlebih dahulu, data latih menjadi seimbang sebelum nilai *Information Gain* dihitung, sehingga perhitungan *entropy* dan *Information Gain* merepresentasikan kedua kelas secara proporsional dan fitur-fitur yang terpilih lebih relevan untuk membedakan kelas asma dan tidak asma. Sebaliknya, ketika *Information Gain* dihitung terlebih dahulu pada data yang masih tidak seimbang (2.268 berbanding 124), nilai *Information Gain* cenderung didominasi oleh kelas mayoritas, sehingga fitur yang dianggap penting lebih merepresentasikan kelas tidak asma, sementara informasi yang berkaitan dengan kelas asma sebagai kelas minoritas kurang terwakili. Akibatnya, meskipun SMOTE diterapkan setelah proses seleksi fitur *Information Gain*, fitur yang telah terpilih kurang mampu menangkap karakteristik kelas asma secara optimal, sehingga kinerja LVQ yang dihasilkan menjadi lebih rendah dibandingkan skenario yang menerapkan SMOTE sebelum *Information Gain*.

Pada penelitian ini, *confusion matrix* digunakan untuk menganalisis model terbaik yang diperoleh dari kombinasi SMOTE, *Information Gain*, dan LVQ. Analisis ini bertujuan untuk mengetahui jumlah data penderita asma yang berhasil teridentifikasi dengan benar serta jumlah data yang mengalami kesalahan klasifikasi. Hasil *confusion matrix* pada Gambar 4. digunakan sebagai dasar dalam perhitungan metrik evaluasi meliputi *precision, recall,* dan *F1-score* yang telah diuraikan sebelumnya.



Gambar 4. *Confusion Matrix* Skenario Terbaik (SMOTE, *Information Gain*, dan LVQ)

Berdasarkan hasil *confusion matrix*, model berhasil mengklasifikasikan 156 data kelas tidak asma dengan benar *(True Negative)* dan 198 data kelas asma dengan benar *(True Positive)*. Sementara itu, terdapat 71 data kelas tidak asma yang salah diklasifikasikan sebagai asma *(False Positive)* dan 29 data kelas asma yang salah diklasifikasikan sebagai tidak asma *(False Negative)*. Hasil pengujian menggunakan confusion matrix ditampilkan pada Gambar 4.

4. KESIMPULAN

Berdasarkan hasil pengujian dan evaluasi seluruh skenario klasifikasi penyakit asma yang telah dilakukan, penerapan SMOTE terbukti menjadi komponen paling penting karena tanpa SMOTE model hanya memprediksi kelas mayoritas dengan *precision*, *recall*, dan *F1-score* bernilai 0% meskipun rata-rata akurasi mencapai 94–95%, sehingga tidak dapat digunakan sebagai alat bantu diagnosis. Skenario terbaik diperoleh pada kombinasi SMOTE, *Information Gain dengan threshold* 0.007, dan LVQ menggunakan fungsi jarak

Euclidean, *learning rate* 0.004, dengan rasio pembagian data 90:10, memperoleh akurasi 77.97%, *precision* 73.61%, *recall* 87.22%, dan *F1-score* 79.84%. Seleksi fitur *Information Gain* dengan *threshold* 0.007 berhasil mengurangi jumlah fitur dari 29 menjadi 22 fitur relevan dan meningkatkan performa model. Selain itu, fungsi jarak *Euclidean* secara konsisten menghasilkan performa yang lebih baik dibandingkan fungsi jarak *Manhattan* dan *Chebyshev*. Keterbatasan akurasi yang diperoleh pada penelitian ini disebabkan oleh persebaran data *class overlap* yang cukup tinggi, dimana data kelas minoritas (asma) tersebar di antara data kelas mayoritas (tidak asma) tanpa batas pemisah yang jelas karena kemiripan karakteristik antara pasien asma dan tidak asma. Penelitian selanjutnya, disarankan untuk menggunakan algoritma klasifikasi yang lebih kompleks atau menerapkan pendekatan *ensemble learning* guna meningkatkan kemampuan model dalam mengenali pola data yang lebih beragam dan kompleks serta memperoleh performa klasifikasi yang lebih baik.

REFERENSI

[1] E. G. Hutahaean, O. Sukma Pratiwi, and R. Afriyani, "Penerapan Metode Forward Chaining Dalam Sistem Pakar Diagnosa Penyakit Asma Menggunakan Bahasa Pemrograman PHP Dan Database Mysql," *Prosiding Seminar Nasional Bisnis Teknologi dan Kesehatan*, no. Vol. 1 No. 1 (2024), Jul. 2024, Accessed: Nov. 12, 2025. [Online]. Available: https://www.ejournal.ummuba.ac.id/index.php/SENABISTEKES/article/view/2436

[2] World Health Organization, "Asthma," https://www.who.int/news-room/fact-sheets/detail/asthma. Accessed: Apr. 15, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/asthma

[3] Risha Justisia Suhendar, Irawan Danismaya, and Kartika Tarwati, "Gambaran Karakteristik Mahasiswa Dengan Asma Bronkial di Universitas Muhammadiyah Sukabumi Tahun 2024," *Jurnal Anestesi*, vol. 2, no. 2, pp. 81–89, Apr. 2024, doi: 10.59680/anestesi.v2i2.1053.

[4] W. Muchsin, *Ensiklopedia Asma: Memahami, Mengelola, dan Hidup Berkualitas dengan Asma*, 1st ed. Purwekerto: PT. Revormasi Jangkar Philosophia, 2025.

[5] S. Aufa, A. Husna, and S. Syahrizal, "Penatalaksanaan Holistik Pasien Anak Dengan Asma Bronkial Melalui Pendekatan Kedokteran Keluarga," *J. Med. Sci.*, vol. 4, no. 2, pp. 127–137, Oct. 2023, doi: 10.55572/jms.v4i2.115.

[6] I. Akbar, F. Supriadi, and D. Indra Junaedi, "Pemanfaatan Machine Learning di Bidang Kesehatan," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, pp. 1744–1749, Jan. 2025, doi: 10.36040/jati.v9i1.12663.

[7] A. U. Dullah, P. Utami, and J. Unjung, "Asthma Classification Using an Adaptive Boosting Model with SVM-SMOTE Sampling," *Journal of Information System Exploration and Research*, vol. 3, no. 1, pp. 1–10, Jan. 2025, doi: 10.52465/joiser.v3i1.486.

[8] Zahab, M. Hussain, and L. S. Parwati, "Prediction of Asthma Disease Using Machine Learning Algorithm," in *Engineering Proceedings*, MDPI AG, Sep. 2025, p. 115. doi: 10.3390/engproc2025107115.

[9] E. A. Sianipar and M. Yasin S, "Optimization of Diabetes Disease Classification Using Learning Vector Quantization Algorithm(LVQ)," *Journal of Computer Science and Informatics Engineering*, vol. 4, no. 2, pp. 72–84, May 2025, doi: 10.55537/cosie.v4i2.1121.

[10] A. Aziz, F. Insani, J. Jasril, and F. Syafria, "Implementasi Metode Learning Vector Quantization (LVQ) Untuk Klasifikasi Keluarga Beresiko Stunting," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3478.

[11] F. Alamri, S. Ningsih, I. Djakaria, D. Wungguli, and I. K. Hasan, "Perbandingan Metode LVQ dan Backpropagation untuk Klasifikasi Status Gizi Anak di Kecamatan Sangkup," *Jurnal Gaussian*, vol. 12, no. 3, pp. 314–321, Sep. 2023, doi: 10.14710/j.gauss.12.3.314-321.

[12] M. Dewi, T. H. Saragih, and R. Herteno, "Penerapan SMOTE-NCL untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Jantung Koroner," *Jurnal Informatika Polinema*, vol. 10, no. 1, Dec. 2023, doi: 10.33795/jip.v10i1.1394.

[13] M. Ibnu Choldun Rachmatullah and S. Armiati, "Menerapkan Smote pada Klasifikasi Data Penyakit Stroke," *Jurnal Ilmiah Manajemen Informatika*, vol. 17, no. 1, 2025, Accessed: Dec. 17, 2025. [Online]. Available: https://ejurnal.ulbi.ac.id/index.php/improve/article/view/4307

[14] A. Syukron, E. Saputro, and P. Widodo, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," 2023. doi: https://doi.org/10.25047/jtit.v10i1.313.

[15] R. A. Azizah, F. Bachtiar, and S. Adinugroho, "Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor dengan Seleksi Fitur Information Gain," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, pp. 605–614, Jun. 2022, doi: 10.25126/jtiik.2022935751.

[16] S. H. Zulaikhah, A. Aziz, and W. Harianto, "Optimasi Algoritma K-Nearest Neighbor (KNN) dengan Normalisasi dan Seleksi Fitur untuk Klasifikasi Penyakit Liver," Sep. 2022. doi: 10.36040/jati.v6i2.4722.

[17] N. Tsawaabul Khair, I. Afrianty, F. Syafria, E. Budianita, and S. Kurnia Gusti, "Penerapan Information Gain Untuk Seleksi Fitur Pada Klasifikasi Jenis Kelamin Tulang Tengkorak Menggunakan Backpropagation," *Media Online)*, vol. 5, no. 4, pp. 666–678, 2025, doi: 10.47065/bulletincsr.v5i4.637.

[18] A. Fitri, I. Afrianty, E. Budianita, and S. Kurnia Gusti, "Implementation of Feature Selection Information Gain in Support Vector Machine Method for Stroke Disease Classification," *Bulletin of Informatics and Data Science*, vol. 4, no. 1, pp. 22–33, 2025, doi: 10.61944/bids.v4i1.116.

[19] F. Meila Azzahra Sofyan, A. Putri Riyandoro, D. Fitriani Maulana, and J. Haerul Jaman, "Penerapan Data Mining dengan Algoritma C5.0 Untuk Prediksi Penyakit Stroke," *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD*, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsk/index

[20] I. Wayan, B. Suryawan, N. Widya Utami, and K. Q. Fredlina, "Analisis Sentimen Review Wisatawan pada Objek Wisata Ubud Menggunakan Algoritma Support Vector Machine," Feb. 2023. doi: https://doi.org/10.51401/jinteks.v5i1.2242.

[21] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1033–1042, Oct. 2024, doi: 10.25126/jtiik.2024117989.

[22] S. Narulita and N. Adi, "Feature Selection Information Gain pada Klasifikasi Pasien Penyakit Jantung (Heart Disease)," *JURMIK (Jurnal Rekam Medis dan Manajemen Informasi Kesehatan)*, vol. 4, no. 1, 2024, doi: 10.53416/jurmik.v4i1.240.

[23] M. Melisa, "Implementasi Learning Vector Quantization (LVQ) Dalam Mengidentifikasi Gula Aren Asli dengan Gula Aren Campuran," *Sci-Tech Journal*, vol. 1, no. 1, pp. 39–51, Jun. 2022, doi: 10.56709/stj.v1i1.18.

[24] F. M. Fathoni, C. A. Putra, and A. L. Nurlaili, "Klasifikasi Penyakit Daun Anggur Menggunakan Metode K-Nearest Neighbor Berdasarkan Gray Level Co-Occurrence Matrix," *Biner : Jurnal Ilmiah Informatika dan Komputer*, vol. 3, no. 1, pp. 8–15, Jan. 2024, doi: 10.32699/biner.v3i1.6332.