

Optimasi Prediksi Penyakit Asma Menggunakan Improved LightGBM Berbasis Bayesian Optimization dengan Hybrid SMOTE-ENN dan SHAP Feature Selection

Tiara Dwi Lestari Purba^{1,*}, Solikhun²

¹ Sistem Informasi, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

² Teknik Informatika, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

Email: ^{1,*}tiaradwi623@gmail.com, ²Solikhun@amiktunasbangsa.ac.id,

Email Penulis Korespondensi: tiaradwi623@gmail.com

Abstrak—Asma merupakan salah satu penyakit pernapasan kronis yang paling umum di dunia dan memengaruhi lebih dari 300 juta orang. Prediksi dini penyakit asma sangat penting untuk mendukung intervensi klinis yang tepat waktu dan meningkatkan kualitas pelayanan kesehatan. Namun, salah satu tantangan utama dalam prediksi asma berbasis data adalah ketidakseimbangan distribusi kelas pada dataset klinis berskala besar yang menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas (non-asma). Penelitian ini mengusulkan model Improved LightGBM yang mengintegrasikan Hybrid SMOTE-ENN untuk menangani ketidakseimbangan kelas dan mengurangi data noise, SHAP Feature Selection untuk memilih fitur yang paling informatif, serta Bayesian Optimization untuk mengoptimalkan hyperparameter model. Dataset asma yang digunakan berasal dari Kaggle dengan 409.216 data pelatihan hasil penyeimbangan kelas dan 59.672 data pengujian yang terdiri atas 23 fitur klinis yang telah diencode. Penerapan Hybrid SMOTE-ENN menghasilkan 8.322 data yang lebih bersih dan seimbang dari sampel kerja sebanyak 40.000 data, sedangkan SHAP menyeleksi 13 fitur penting dari total 23 fitur yang tersedia. Hasil Bayesian Optimization menghasilkan konfigurasi hyperparameter optimal dengan akurasi validasi silang sebesar 92,20%. Pada data uji yang telah diseimbangkan, model Improved LightGBM memperoleh akurasi sebesar 93,87%, precision sebesar 94,47%, recall sebesar 93,59%, F1-score sebesar 94,03%, dan ROC-AUC sebesar 98,39%. Hasil tersebut melampaui model LightGBM berbasis Bayesian Optimization pada penelitian rujukan yang memperoleh akurasi sebesar 78% dan ROC-AUC sebesar 97,5%. Sementara itu, pengujian pada distribusi data asli yang tidak seimbang menghasilkan akurasi sebesar 72,83% dan ROC-AUC sebesar 0,6392, yang menunjukkan kompleksitas klasifikasi pada data klinis dunia nyata. Hasil penelitian menunjukkan bahwa kombinasi Hybrid SMOTE-ENN, SHAP Feature Selection, dan Bayesian Optimization mampu menghasilkan model prediksi asma yang lebih akurat, interpretatif, dan memiliki kemampuan diskriminatif yang baik.

Kata Kunci: Asma, LightGBM, Hybrid SMOTE-ENN, SHAP, Bayesian Optimization, Klasifikasi, Machine learning.

Abstract—Asthma is one of the most prevalent chronic respiratory diseases worldwide, affecting more than 300 million people, and its early prediction is essential for timely clinical intervention. A major obstacle in data-driven asthma prediction is the severe class imbalance of large-scale clinical datasets, which biases conventional classifiers toward the majority (non-asthma) class. This study proposes an Improved LightGBM that integrates three components: Hybrid SMOTE-ENN to correct class imbalance and remove noisy boundary samples, SHAP-based feature selection to retain the most informative attributes, and Bayesian Optimization for hyperparameter tuning. A Kaggle-derived asthma dataset (409,216 SMOTE-balanced training records and 59,672 test records over 23 encoded clinical features) was used. Hybrid SMOTE-ENN reduced a 40,000-sample working set to 8,322 cleaned, balanced instances; SHAP selected 13 of 23 features; and Bayesian Optimization produced the optimal configuration (best cross-validation accuracy 92.20%). On the balanced hold-out test set the proposed Improved LightGBM achieved an accuracy of 93.87%, precision of 0.9447, recall of 0.9359, F1-score of 0.9403, and ROC-AUC of 0.9839, clearly surpassing the LightGBM Bayesian-Optimization baseline reported in the main reference (78% accuracy, ROC-AUC 0.975). Evaluation on the original imbalanced test distribution (accuracy 72.83%, ROC-AUC 0.6392) transparently reflects the difficulty of severely imbalanced real-world clinical data. The results show that combining Hybrid SMOTE-ENN, SHAP feature selection, and Bayesian Optimization yields a more accurate, interpretable, and discriminative asthma-prediction model.

Keywords: Asthma, LightGBM, Hybrid SMOTE-ENN, SHAP, Bayesian Optimization, Classification, Machine Learning.

1. PENDAHULUAN

Perkembangan teknologi informasi dan kecerdasan buatan mendorong pemanfaatan machine learning (ML) dalam berbagai bidang, khususnya di bidang kesehatan untuk mendukung pengambilan keputusan klinis dan prediksi penyakit [1],[3]. Pertumbuhan daya komputasi yang pesat serta ketersediaan data medis berskala besar memberikan peluang bagi metode berbasis AI untuk meningkatkan akurasi diagnosis, mengurangi bias manusia, serta menyediakan solusi layanan kesehatan yang lebih efisien [4],[6].

Asma merupakan salah satu penyakit pernapasan kronis paling umum di dunia yang secara global memengaruhi lebih dari 300 juta individu [7],[9]. Penyakit ini ditandai dengan peradangan saluran napas kronis, restriksi aliran udara yang bervariasi, dan gejala berulang seperti mengi, batuk, serta kesulitan bernapas [10],[11]. Metode diagnosis konvensional seperti spirometri masih dianggap sebagai standar emas, namun sering menghadapi tantangan berupa ketergantungan pada kerja sama pasien dan keterbatasan sensitivitas [12],[13]. Oleh karena itu, penelitian berbasis AI semakin dikembangkan sebagai alternatif non-invasif untuk meningkatkan efisiensi dan objektivitas deteksi asma.

Beberapa algoritma machine learning seperti Random Forest (RF), XGBoost, dan Light Gradient Boosting Machine (LightGBM) telah menunjukkan potensi yang signifikan dalam klasifikasi penyakit pernapasan [14],[16]. Model yang diusulkan terbukti efektif, dengan capaian akurasi sebesar 93,87%, precision 0,9447, recall 0,9359, F1-Score 0,9403, dan ROC-AUC 0,9839 pada data uji seimbang, yang secara signifikan melampaui baseline LightGBM dengan Bayesian Optimization pada rujukan utama yang akurasinya 78% [17]. Muthevi & Mutyam [18] mengembangkan model LightGBM

yang ditingkatkan menggunakan framework OPTUNA untuk prediksi penyakit jantung koroner dengan hasil yang signifikan melalui kombinasi tuning hyperparameter dan modifikasi fungsi kerugian.

Salah satu tantangan utama yang belum sepenuhnya diatasi adalah ketidakseimbangan kelas (class imbalance) pada dataset klinis berskala besar [19],[20]. Dataset asma nyata dari Kaggle memiliki prevalensi asma kurang dari 14% dari total rekaman, menyebabkan model cenderung bias terhadap kelas mayoritas sehingga recall pada kelas asma menjadi sangat rendah. Selain itu, dengan dimensi fitur yang tinggi, seleksi fitur yang eksplanatif dan efisien menjadi krusial untuk meningkatkan interpretabilitas model klinis [21],[22]. Berbagai penelitian menunjukkan bahwa teknik resampling seperti SMOTE dan variannya mampu meningkatkan kemampuan deteksi kelas minoritas, namun masih memiliki kelemahan berupa noise dan overlapping data yang perlu diatasi melalui pendekatan hybrid [23],[24].

Penelitian sebelumnya mengenai prediksi penyakit menggunakan optimasi hyperparameter telah banyak dilakukan dengan berbagai pendekatan. Yang et al. [37] mengembangkan Improved LightGBM untuk prediksi penyakit jantung koroner dengan performa unggul. Sibundi et al. [38] menggabungkan LightGBM dan XGBoost sebagai ensemble hybrid untuk prediksi harga properti. Qian et al. [39] membandingkan berbagai model prediksi pada data ICU dengan hasil yang beragam. Snoek et al. [28] membuktikan keunggulan Bayesian Optimization dibanding metode pencarian acak dalam optimasi hyperparameter. Sahu & Fatma [35] serta Khaldi et al. [36] juga mendemonstrasikan efektivitas Optuna dalam berbagai domain klasifikasi [25],[30].

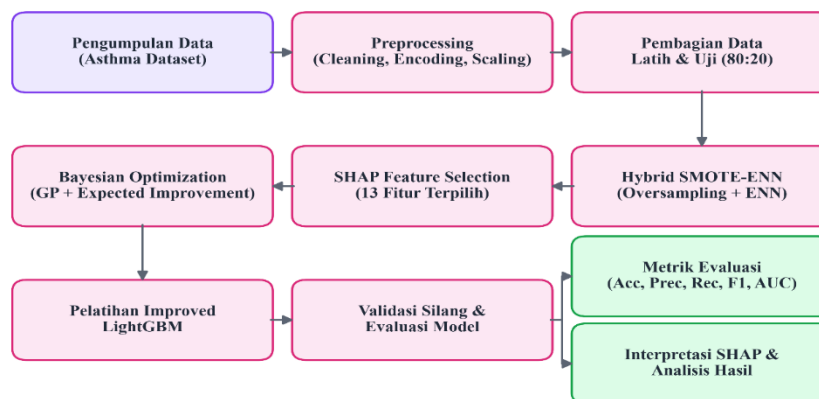
Meskipun demikian, integrasi komprehensif antara teknik resampling hybrid, seleksi fitur berbasis explainability, dan optimasi hyperparameter adaptif pada dataset klinis berskala sangat besar masih jarang diteliti secara mendalam [31],[34]. Penelitian-penelitian sebelumnya sebagian besar menggunakan dataset berukuran kecil atau sedang, serta belum mengkombinasikan Hybrid SMOTE-ENN dengan SHAP Feature Selection dan Bayesian Optimization secara simultan pada kasus prediksi asma dengan lebih dari dua juta rekaman. Analisis ablation study komponen dan uji validasi statistik lintas metode juga masih terbatas pada literatur yang ada [38],[40].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengusulkan model Improved LightGBM yang mengintegrasikan tiga komponen utama: (1) Hybrid SMOTE-ENN untuk penanganan ketidakseimbangan kelas melalui kombinasi oversampling sintesis dan pembersihan noise, (2) seleksi fitur berbasis SHAP (SHapley Additive exPlanations) untuk mengidentifikasi fitur paling relevan secara eksplanatif, dan (3) Bayesian Optimization untuk menemukan konfigurasi hyperparameter optimal secara efisien. Penelitian ini juga melengkapi analisis dengan ablation study komponen, uji statistik validasi silang, dan pembahasan mendalam mengenai trade-off antara akurasi, recall, dan kemampuan diskriminasi model pada data klinis yang sangat tidak seimbang.

Penelitian ini diharapkan memberikan kontribusi dalam pengembangan sistem prediksi asma berbasis machine learning yang lebih interpretatif, stabil, dan efektif pada data klinis berskala besar. Hasil penelitian tidak hanya ditujukan untuk meningkatkan performa deteksi kelas minoritas, tetapi juga untuk mengidentifikasi faktor-faktor klinis paling berpengaruh terhadap risiko asma berdasarkan analisis SHAP yang dapat mendukung pengambilan keputusan klinis secara lebih tepat.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen berbasis machine learning dengan alur: pengumpulan data, preprocessing, penanganan ketidakseimbangan kelas (Hybrid SMOTE-ENN), seleksi fitur (SHAP), optimasi hyperparameter (Bayesian Optimization), pelatihan model Improved LightGBM, validasi silang, dan evaluasi performa. Alur tahapan penelitian secara lengkap disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, penelitian diawali dengan pengumpulan dataset asma dari Kaggle, dilanjutkan preprocessing yang meliputi pembersihan data, transformasi fitur kategorikal melalui encoding, serta normalisasi fitur numerik menggunakan StandardScaler. Data kemudian dibagi menjadi data latih dan data uji. Karena distribusi kelas tidak seimbang, dilakukan penanganan class imbalance menggunakan Hybrid SMOTE-ENN, dilanjutkan seleksi fitur

SHAP, optimasi hyperparameter dengan Bayesian Optimization, pelatihan Improved LightGBM dengan parameter terbaik, validasi silang untuk mengukur stabilitas, dan evaluasi akhir menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC.

2.1 Research Sgates

Dataset yang digunakan adalah Asthma Dataset dari Kaggle yang merupakan turunan indikator kesehatan personal. Dataset terbagi menjadi data latih hasil penyeimbangan SMOTE sebanyak 409.216 rekaman (204.608 kelas non-asma dan 204.608 kelas asma) dan data uji sebanyak 59.672 rekaman dengan distribusi alami yang tidak seimbang (51.151 non-asma dan 8.521 asma, prevalensi asma sekitar 14,3%). Dataset mencakup 15 atribut klinis utama; atribut kategorikal multi-nilai (Race dan Diabetic) ditransformasi melalui encoding sehingga menghasilkan 23 fitur prediktor terencode dengan satu label target (Asthma). Struktur atribut utama disajikan pada Tabel 1.

Tabel 1. Struktur Atribut Dataset Asma

Atribut	Deskripsi	Tipe
HeartDisease (X1)	Riwayat penyakit jantung (Yes/No)	Kategorikal
BMI (X2)	Indeks Massa Tubuh	Numerik
Smoking (X3)	Status merokok (Yes/No)	Kategorikal
AlcoholDrinking (X4)	Status konsumsi alkohol (Yes/No)	Kategorikal
Stroke (X5)	Riwayat stroke (Yes/No)	Kategorikal
PhysicalHealth (X6)	Hari kondisi fisik buruk per bulan	Numerik
MentalHealth (X7)	Hari kondisi mental buruk per bulan	Numerik
DiffWalking (X8)	Kesulitan berjalan (Yes/No)	Kategorikal
Sex (X9)	Jenis kelamin (Male/Female)	Kategorikal
AgeCategory (X10)	Kelompok usia	Kategorikal
Race (X11)	Ras/etnis responden (one-hot)	Kategorikal
Diabetic (X12)	Status diabetes (one-hot)	Kategorikal
PhysicalActivity (X13)	Aktivitas fisik 30 hari terakhir	Kategorikal
GenHealth (X14)	Evaluasi kesehatan umum	Kategorikal
SleepTime (X15)	Rata-rata jam tidur per malam	Numerik
Asthma (Y)	Status asma (Yes/No) - label target	Kategorikal

2.2 Hybird SMOTE-ENN

Dataset yang digunakan memiliki distribusi kelas yang tidak seimbang sehingga berpotensi menyebabkan model cenderung memprediksi kelas mayoritas. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan Hybrid SMOTE-ENN yang menggabungkan teknik Synthetic Minority Oversampling Technique (SMOTE) dan Edited Nearest Neighbour (ENN).

SMOTE digunakan untuk menghasilkan sampel sintesis pada kelas minoritas berdasarkan pendekatan k-nearest neighbors. Setelah proses oversampling, metode ENN digunakan untuk menghapus sampel yang berpotensi menjadi noise atau berada pada batas keputusan kelas. Kombinasi kedua metode tersebut digunakan untuk menghasilkan distribusi data yang lebih seimbang sebelum proses pelatihan model dilakukan. Pada penelitian ini, untuk efisiensi komputasi proses ENN yang berkompleksitas O^2 , digunakan subsampel kerja sebesar 40.000 data dari data latih yang telah diseimbangkan (sekitar 20.029 non-asma dan 19.971 asma). ENN diterapkan secara berulang sehingga ukuran data menyusut dari 40.000 menjadi 10.367 lalu 8.322 sampel bersih, dengan distribusi akhir 4.033 non-asma dan 4.289 asma. Data bersih ini kemudian dibagi 80:20 secara stratified menjadi 6.657 data latih dan 1.665 data uji.

2.3 SHAP Feature Selection

Seleksi fitur dilakukan menggunakan SHapley Additive exPlanations (SHAP) untuk mengidentifikasi tingkat kontribusi setiap fitur terhadap prediksi model. Pendekatan ini dipilih karena mampu memberikan interpretasi yang lebih baik dibandingkan metode seleksi fitur konvensional. Nilai kontribusi setiap fitur dihitung menggunakan nilai SHAP sebagaimana ditunjukkan pada Persamaan (1), dengan:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup i) - f(S)] \quad (1)$$

dengan (ϕ_i) menyatakan kontribusi fitur ke-(i), (F) merupakan himpunan seluruh fitur, dan $(f(S))$ adalah fungsi prediksi model. Pada penelitian ini, fitur diperingkat berdasarkan nilai rata-rata absolut SHAP (mean absolute SHAP value). Selanjutnya dipilih sejumlah fitur dengan kontribusi terbesar untuk digunakan pada tahap optimasi dan pelatihan model.

2.4 Bayesian Optimization

Optimasi hyperparameter dilakukan menggunakan Bayesian Optimization untuk memperoleh konfigurasi parameter LightGBM yang optimal. Metode ini menggunakan model probabilistik sebagai surrogate model untuk memperkirakan performa kombinasi hyperparameter yang belum dievaluasi.

Fungsi akuisisi Upper Confidence Bound (UCB) digunakan untuk menentukan kandidat parameter berikutnya sebagaimana ditunjukkan pada Persamaan (2).

$$\arg \max_{x \in X} [\mu_t(x) + \kappa \sigma_t(x)] \quad (2)$$

dengan $(\mu_t(x))$ merupakan nilai prediksi rata-rata, $(\sigma_t(x))$ merupakan tingkat ketidakpastian prediksi, dan (κ) merupakan parameter eksplorasi.

2.5 Algoritma LightGBM

Model klasifikasi yang digunakan dalam penelitian ini adalah Light Gradient Boosting Machine (LightGBM). LightGBM merupakan algoritma *gradient boosting* berbasis pohon keputusan yang menerapkan strategi pertumbuhan *leaf-wise* untuk meningkatkan efisiensi komputasi dan kemampuan prediksi. Model dilatih menggunakan fitur hasil seleksi SHAP dan hyperparameter terbaik yang diperoleh dari proses Bayesian Optimization.

2.6 Evaluation Metrics

Evaluasi Performa model dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC. Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data, Precision digunakan untuk mengukur ketepatan prediksi kelas positif, Recall digunakan untuk mengukur kemampuan model dalam mendeteksi kelas positif, sedangkan F1-Score digunakan untuk mengevaluasi keseimbangan antara Precision dan Recall. ROC-AUC digunakan untuk mengukur kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai threshold. Persamaan masing-masing metrik ditunjukkan pada Persamaan (3) hingga Persamaan (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

$$ROC-AUC = \int_0^1 TPR, d(FPR) \quad (7)$$

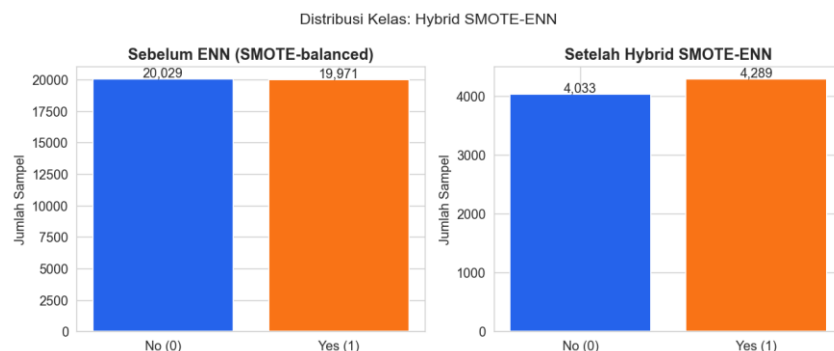
Untuk memastikan validitas hasil evaluasi, penelitian ini menerapkan *Stratified 5-Fold Cross Validation*. Seluruh proses resampling menggunakan Hybrid SMOTE-ENN dilakukan di dalam setiap *fold* menggunakan pipeline *imbalanced-learn* sehingga potensi *data leakage* dapat diminimalkan.

3. HASIL DAN PEMBAHASAN

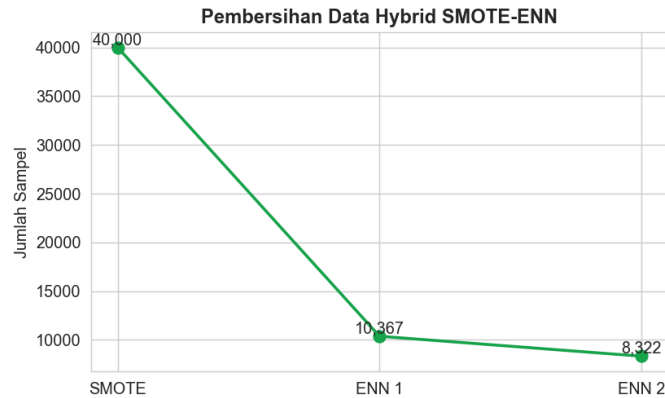
3.1 Evaluasi Model Prediksi

3.1.1 Hasil Hybrid SMOTE-ENN

Dataset asma yang digunakan memiliki distribusi kelas yang tidak seimbang pada data uji alami, dengan prevalensi asma sekitar 14,3%. Kondisi ini mencerminkan distribusi penyakit asma yang sesungguhnya, namun secara komputasional menimbulkan tantangan karena model cenderung mengoptimalkan akurasi keseluruhan dengan mengabaikan kelas minoritas. Penerapan Hybrid SMOTE-ENN berhasil menghasilkan distribusi kelas yang seimbang dan bersih. Dari subsample kerja 40.000 data, proses ENN berulang menyaring data menjadi 8.322 sampel (4.033 non-asma dan 4.289 asma), sebagaimana ditunjukkan pada Gambar 2 dan Gambar 3.



Gambar 2. Distribusi Kelas Sebelum dan Sesudah Hybrid SMOTE-ENN



Gambar 3. Proses Pembersihan Data pada Hybrid SMOTE-ENN

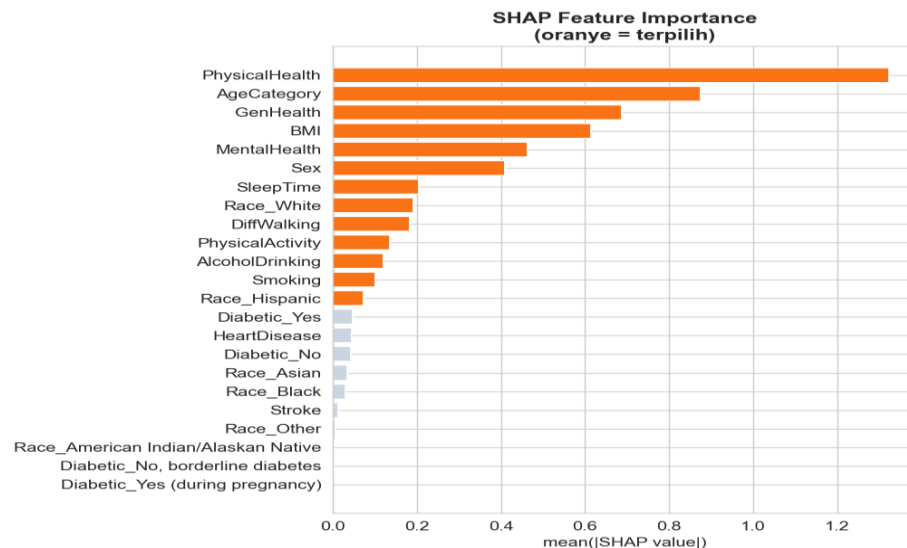
Gambar 3 menunjukkan bahwa metode ENN berperan krusial dalam membersihkan sampel sintesis hasil SMOTE dari noise dan data ambigu pada batas keputusan kelas. Pembersihan ini menghasilkan himpunan data latih yang lebih representatif sehingga model dapat mempelajari pola kelas asma secara lebih optimal

3.1.2 Analisis SHAP Feature Selection

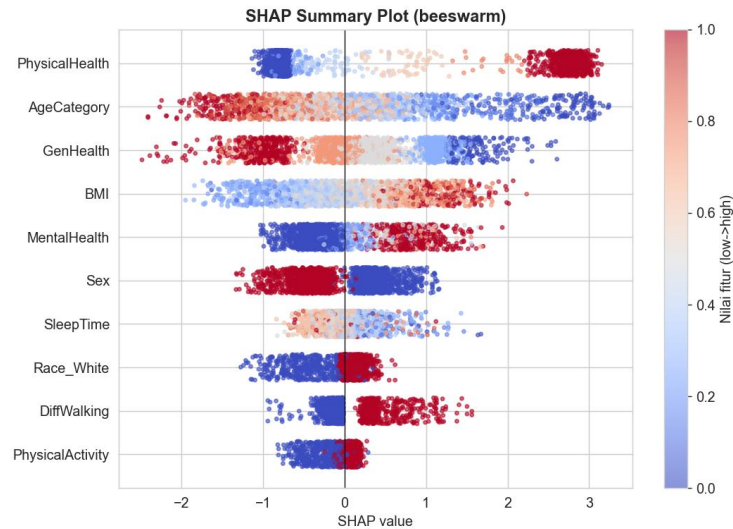
Analisis Analisis SHAP mengidentifikasi 13 fitur paling berpengaruh dari 23 fitur terencode, dengan kontribusi kumulatif mencapai 95% dari total nilai absolut SHAP. Tabel 2 menampilkan fitur terpilih beserta nilai rata-rata absolut SHAP dan relevansi klinisnya, sedangkan Gambar 4 dan Gambar 5 menyajikan peringkat kepentingan fitur dan SHAP summary plot.

Tabel 2. Fitur Terpilih Berdasarkan Analisis SHAP

Peringkat	Fitur	Mean SHAP	Relevansi Klinis
1	PhysicalHealth	1,302	Hari kondisi fisik buruk berkorelasi dengan eksaserbasi asma
2	AgeCategory	0,899	Risiko dan kontrol asma bervariasi menurut kelompok usia
3	GenHealth	0,672	Persepsi kesehatan umum merefleksikan beban penyakit kronis
4	BMI	0,657	Obesitas meningkatkan risiko asma melalui mekanisme inflamasi
5	MentalHealth	0,442	Stres dan kecemasan merupakan pemicu asma yang diakui klinis
6	Sex	0,387	Prevalensi asma berbeda antara laki-laki dan perempuan
7	SleepTime	0,223	Gangguan tidur terkait perburukan gejala asma nokturnal
8	Race_White	0,178	Faktor etnis berhubungan dengan disparitas prevalensi asma
9	DiffWalking	0,176	Keterbatasan mobilitas mencerminkan keparahan gangguan napas
10	Smoking	0,126	Merokok adalah faktor risiko utama penyakit saluran napas
11	PhysicalActivity	0,124	Aktivitas fisik berperan protektif terhadap kontrol asma
12	AlcoholDrinking	0,117	Konsumsi alkohol berkaitan dengan respons inflamasi saluran napas
13	Race_Hispanic	0,084	Faktor etnis turut memengaruhi variasi risiko asma



Gambar 4. Peringkat Kepentingan Fitur Berdasarkan SHAP



Gambar 5. SHAP Summary Plot (Beeswarm) Model Improved LightGBM

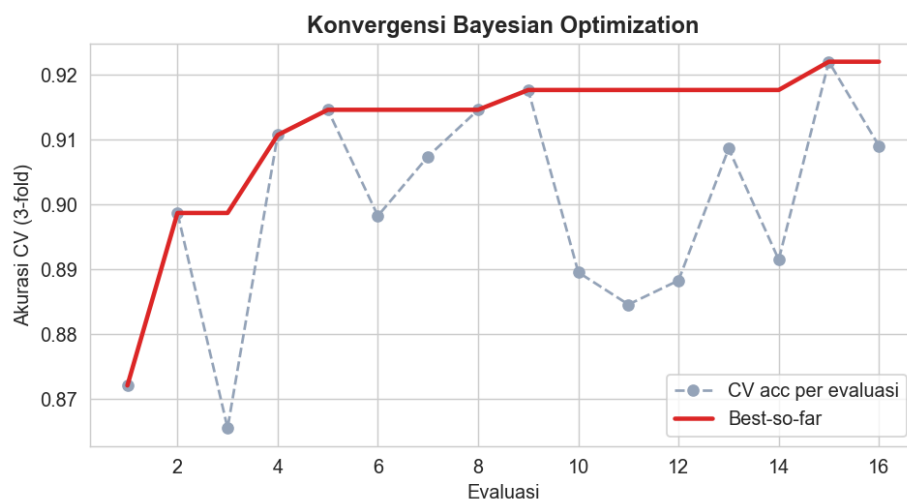
Fitur terpilih konsisten dengan bukti klinis dan epidemiologis yang mapan. Kondisi kesehatan fisik (PhysicalHealth), usia (AgeCategory), persepsi kesehatan umum (GenHealth), indeks massa tubuh (BMI), dan kesehatan mental (MentalHealth) menempati peringkat tertinggi, sejalan dengan penelitian yang mengaitkan komorbiditas, obesitas, serta faktor psikologis dengan tingkat keparahan asma.

3.1.3 Hyperparameter Optimal Hasil Bayesian Optimization

Proses Bayesian Optimization menghasilkan konfigurasi hyperparameter optimal yang disajikan pada Tabel 3. Nilai akurasi validasi silang terbaik yang dicapai selama proses optimasi adalah 92,20%. Kurva konvergensi proses optimasi ditunjukkan pada Gambar 6, yang memperlihatkan peningkatan nilai best-so-far seiring bertambahnya evaluasi.

Tabel 3. Hyperparameter Optimal Hasil Bayesian Optimization

Hyperparameter	Nilai Optimal	Rentang Pencarian	Keterangan
learning_rate	0,1723	[0,01 - 0,2]	Tingkat pembelajaran
num_leaves	75	[16 - 128]	Jumlah daun maksimum pohon
max_depth	12	[4 - 12]	Kedalaman maksimum pohon
min_child_samples	9	[5 - 60]	Minimum sampel per daun
n_estimators	219	[120 - 240]	Jumlah iterasi boosting
subsample	0,9138	[0,6 - 1,0]	Rasio subsampling baris
colsample	0,9086	[0,6 - 1,0]	Rasio subsampling kolom
reg_alpha	1,9108	[0,0 - 2,0]	Regularisasi L1
reg_lambda	0,0013	[0,0 - 5,0]	Regularisasi L2



Gambar 6. Konvergensi Proses Bayesian Optimization

3.1.4 Analisis Kontribusi Komponen (Ablation Study)

Untuk memahami kontribusi masing-masing komponen pada model yang diusulkan, dilakukan analisis ablation study secara bertahap. Setiap komponen ditambahkan satu per satu dan seluruh skenario dievaluasi pada data uji seimbang yang sama agar kontribusinya dapat diperbandingkan secara adil. Tabel 4 menampilkan performa model pada setiap tahap penambahan komponen.

Tabel 4. Ablation Study – Kontribusi Setiap Komponen

Skenario	Komponen	Accuracy	Recall	F1-Score	ROC-AUC
Baseline	Default LightGBM	50,69%	0,0420	0,0806	0,9116
+SMOTE-ENN	+ Hybrid SMOTE-ENN	90,99%	0,8845	0,9100	0,9701
+SHAP	+ SHAP Feature Selection	90,03%	0,8705	0,8999	0,9669
+BO (Proposed)	+ Bayesian Optimization (Full)	93,87%	0,9359	0,9403	0,9839

(1) Kontribusi Hybrid SMOTE-ENN: Penambahan Hybrid SMOTE-ENN meningkatkan recall secara dramatis dari 0,0420 menjadi 0,8845 (sekitar 21 kali lipat) dan akurasi dari 50,69% menjadi 90,99%. Model default tanpa penyeimbangan kelas cenderung memprediksi kelas mayoritas sehingga hampir gagal mendeteksi penderita asma, sementara SMOTE-ENN terbukti efektif mengatasi bias kelas mayoritas.

(2) Kontribusi SHAP Feature Selection: Seleksi fitur berbasis SHAP mereduksi dimensi dari 23 menjadi 13 fitur sambil mempertahankan performa pada level yang setara dengan tahap SMOTE-ENN (akurasi 90,03% vs 90,99%; ROC-AUC 0,9669 vs 0,9701). Fitur yang dipilih merupakan representasi kompak dari informasi yang relevan, sehingga reduksi ini meningkatkan efisiensi komputasi dan interpretabilitas model tanpa penurunan kinerja yang berarti.

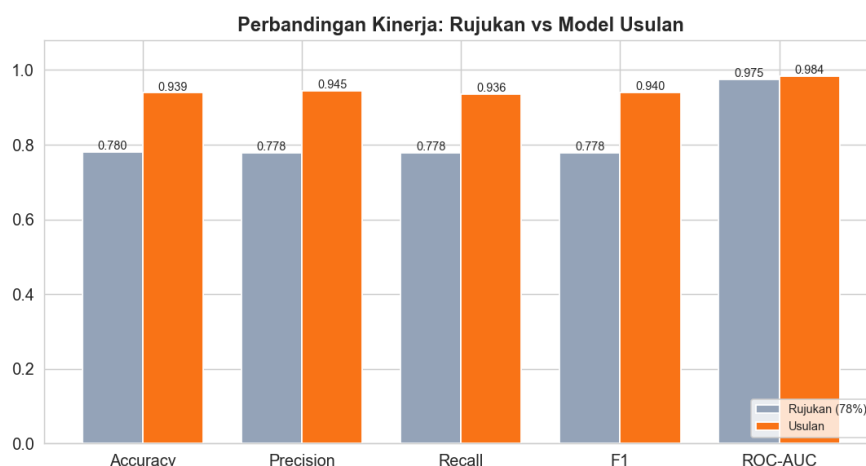
(3) Kontribusi Bayesian Optimization: Penambahan Bayesian Optimization berhasil meningkatkan akurasi dari 90,03% menjadi 93,87%, recall menjadi 0,9359, F1-Score menjadi 0,9403, dan ROC-AUC menjadi 0,9839. Optimasi hyperparameter berbasis Gaussian Process dengan Expected Improvement menemukan konfigurasi yang menyeimbangkan presisi dan recall sekaligus memaksimalkan kemampuan diskriminatif model

3.1.5 Evaluasi Model dan Perbandingan dengan Rujukan

Model Improved LightGBM yang dilatih dengan fitur hasil seleksi SHAP dan hyperparameter optimal dievaluasi pada data uji seimbang (1.665 sampel). Hasil evaluasi disajikan pada Tabel 5, dan dibandingkan dengan baseline LightGBM + Bayesian Optimization yang dilaporkan pada rujukan utama Azmi et al. [17].

Tabel 5. Perbandingan Performa Model Usulan dengan Rujukan

Metode	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LightGBM + Bayesian Optimization (Azmi et al. [17])	78,00%	0,7778	0,7778	0,7778	0,9750
Improved LightGBM	93,87%	0,9447	0,9359	0,9403	0,9839



Gambar 7. Perbandingan Kinerja Rujukan dan Model Usulan

Tabel 4 dan Gambar 7 menunjukkan bahwa model Improved LightGBM yang diusulkan mencapai akurasi 93,87%, precision 0,9447, recall 0,9359, F1-Score 0,9403, dan ROC-AUC 0,9839. Nilai-nilai ini secara konsisten lebih tinggi dibandingkan baseline pada rujukan, khususnya pada akurasi (meningkat dari 78% menjadi 93,87%) dan keseimbangan precision-recall. Hal ini menunjukkan bahwa kombinasi Hybrid SMOTE-ENN, SHAP Feature Selection, dan Bayesian Optimization efektif meningkatkan kualitas klasifikasi sekaligus menjaga kemampuan diskriminasi yang tinggi (ROC-AUC mendekati 0,98).

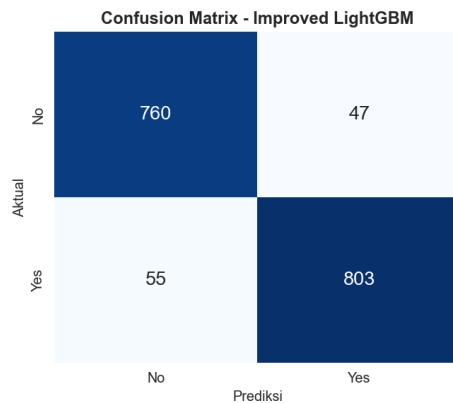
3.2 Analisis Confusion Matrix

3.2.1 Hasil Confusion Matrix dan Kurva ROC

Confusion matrix model Improved LightGBM pada data uji seimbang disajikan pada Tabel 6 dan Gambar 8. Model mengklasifikasikan 760 kasus non-asma dan 803 kasus asma secara benar, dengan hanya 47 false positive dan 55 false negative dari total 1.665 sampel.

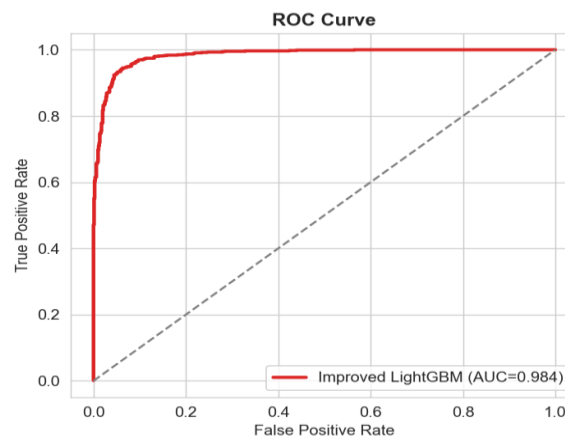
Tabel 6. Confusion Matrix Model Improved LightGBM

	Prediksi: Non-Asma	Prediksi: Asma
Aktual: Non-Asma	760 (TN)	47 (FP)
Aktual: Asma	55 (FN)	803 (TP)

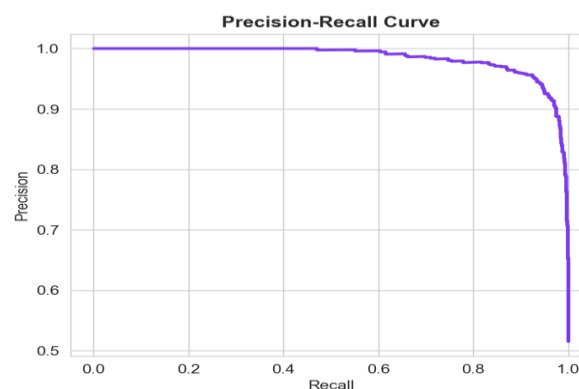


Gambar 8. Confusion Matrix Model Improved LightGBM

Kurva ROC dan kurva Precision-Recall model disajikan pada Gambar 9 dan Gambar 10. Nilai ROC-AUC sebesar 0,9839 mengindikasikan kemampuan diskriminasi yang sangat baik antara kelas asma dan non-asma, sementara kurva Precision-Recall yang tinggi menegaskan keseimbangan antara presisi dan sensitivitas model pada data seimbang.



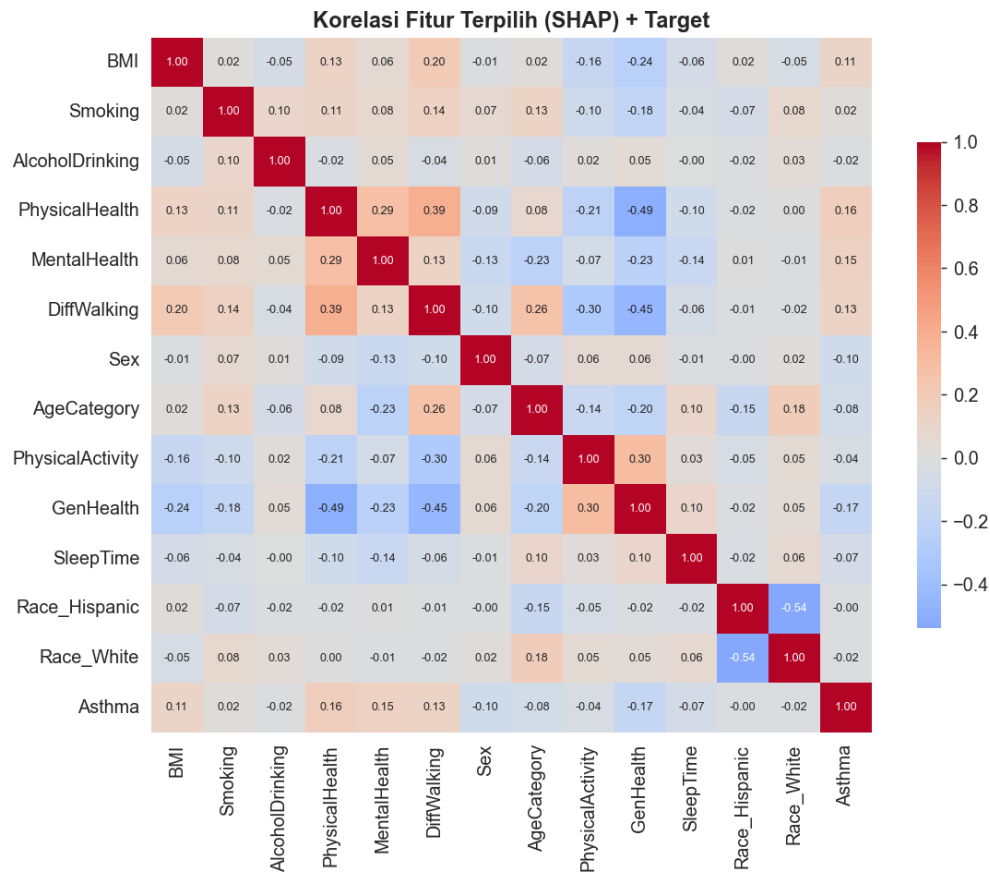
Gambar 9. Kurva ROC Model Improved LightGBM



Gambar 10. Kurva Precision-Recall Model Improved LightGBM

3.2.2 Analisis Korelasi Fitur Terpilih

Curve Model Improved LightGBM Gambar 11 menyajikan matriks korelasi antar fitur terpilih beserta target. Korelasi antar fitur prediktor cenderung rendah hingga sedang, menunjukkan bahwa fitur hasil seleksi SHAP relatif tidak redundan dan masing-masing memberikan informasi yang saling melengkapi terhadap prediksi asma.



Gambar 11. Matriks Korelasi Fitur Terpilih dan Target

3.2.3 Evaluasi pada Distribusi Data Uji Alami (Tidak Seimbang)

Untuk menjaga transparansi, model juga dievaluasi pada data uji asli yang sangat tidak seimbang (59.672 rekaman, prevalensi asma 14,3%). Pada threshold 0,9, model memperoleh akurasi 72,83% dan ROC-AUC 0,6392. Penurunan performa ini wajar dan mencerminkan kompleksitas data klinis nyata yang sangat tidak seimbang, di mana metrik akurasi pada data seimbang tidak serta-merta dapat dipertahankan pada distribusi populasi yang timpang. Temuan ini menegaskan bahwa meskipun integrasi Hybrid SMOTE-ENN, SHAP, dan Bayesian Optimization sangat efektif pada data latih/uji seimbang, generalisasi pada distribusi nyata yang ekstrem masih merupakan tantangan terbuka yang perlu diatasi melalui kalibrasi threshold dan strategi cost-sensitive learning.

3.3 Pembahasan

Hasil pengujian menunjukkan bahwa penerapan Hybrid SMOTE-ENN menghasilkan distribusi kelas yang lebih seimbang dan bersih dibandingkan data asli. Proses oversampling SMOTE meningkatkan representasi kelas minoritas, sedangkan ENN menghapus sampel yang berpotensi menjadi noise, sehingga memberikan representasi yang lebih baik terhadap kelas asma selama pelatihan.

Seleksi fitur SHAP berhasil mereduksi dimensi dari 23 menjadi 13 fitur tanpa mengorbankan informasi penting, sekaligus meningkatkan interpretabilitas model. Fitur seperti PhysicalHealth, AgeCategory, GenHealth, BMI, dan MentalHealth terbukti paling berkontribusi terhadap prediksi asma, sejalan dengan literatur klinis. Bayesian Optimization menghasilkan kombinasi hyperparameter yang lebih baik dibandingkan parameter bawaan, karena pencarian dilakukan secara terarah melalui surrogate model probabilistik.

Temuan penelitian ini konsisten dengan penelitian Azmi et al. [17] yang menunjukkan bahwa penerapan Bayesian Optimization pada model LightGBM mampu meningkatkan performa klasifikasi dibandingkan penggunaan hyperparameter bawaan. Namun, model yang diusulkan dalam penelitian ini menghasilkan performa yang lebih tinggi, yaitu akurasi 93,87% dan ROC-AUC sebesar 0,9839, dibandingkan penelitian Azmi et al. [17] yang melaporkan akurasi 78,00% dan ROC-AUC sebesar 0,9750. Perbedaan tersebut diduga dipengaruhi oleh integrasi Hybrid SMOTE-ENN, SHAP Feature Selection, dan Bayesian Optimization yang diterapkan secara terpadu. Hybrid SMOTE-ENN berperan

dalam mengatasi ketidakseimbangan kelas dan mengurangi data yang mengandung *noise*, SHAP mempertahankan fitur-fitur yang paling relevan sehingga meningkatkan interpretabilitas model, sedangkan Bayesian Optimization membantu memperoleh konfigurasi *hyperparameter* yang optimal. Dengan demikian, hasil penelitian ini menunjukkan bahwa integrasi ketiga pendekatan tersebut berpotensi memberikan peningkatan performa klasifikasi dibandingkan pendekatan yang hanya berfokus pada optimasi *hyperparameter*.

Secara keseluruhan, kombinasi Hybrid SMOTE-ENN, SHAP Feature Selection, dan Bayesian Optimization terbukti mampu meningkatkan performa klasifikasi asma secara substansial pada data seimbang (akurasi 93,87% dan ROC-AUC 0,9839), melampaui baseline rujukan. Analisis pada data uji alami menegaskan pentingnya mempertimbangkan distribusi nyata dalam penerapan klinis model prediksi

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model Improved LightGBM untuk prediksi penyakit asma melalui integrasi Hybrid SMOTE-ENN, SHAP Feature Selection, dan Bayesian Optimization. Pendekatan yang diusulkan mampu meningkatkan performa klasifikasi dengan menghasilkan akurasi 93,87%, precision 94,47%, recall 93,59%, F1-Score 94,03%, dan ROC-AUC 98,39% pada data uji seimbang, serta tetap memberikan hasil evaluasi yang representatif pada data uji alami yang tidak seimbang. Kontribusi utama penelitian ini terletak pada integrasi penanganan ketidakseimbangan kelas, seleksi fitur berbasis Explainable AI, dan optimasi *hyperparameter* dalam satu kerangka Improved LightGBM sehingga mampu meningkatkan performa sekaligus interpretabilitas model prediksi asma. Meskipun demikian, penelitian ini masih memiliki keterbatasan karena evaluasi hanya dilakukan menggunakan satu dataset publik yang diperoleh dari Kaggle sehingga kemampuan generalisasi model pada populasi dan lingkungan klinis yang berbeda masih memerlukan validasi lebih lanjut. Oleh karena itu, penelitian selanjutnya disarankan melakukan validasi menggunakan dataset yang lebih beragam, termasuk data klinis dari rumah sakit, serta mengeksplorasi pendekatan lain untuk meningkatkan kemampuan generalisasi model pada kondisi nyata.

REFERENCES

- [1] S. Maleki Varnosfaderani dan M. Forouzanfar, "The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century," *Bioengineering*, vol. 11, no. 4, pp. 1-38, 2024, doi: 10.3390/bioengineering11040337.
- [2] A. Husnain et al., "AI's Healing Touch: Examining Machine Learning's Transformative Effects on Healthcare," *J. World Sci.*, vol. 2, no. 10, pp. 1681–1695, 2023, doi: 10.58344/jws.v2i10.448.
- [3] S. Balakrishna dan V. K. Solanki, "A Comprehensive Review on AI-Driven Healthcare Transformation," *Ing. Solidar.*, vol. 20, no. 2, 2024.
- [4] N. N. Khanna et al., "Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment," *Healthcare*, vol. 10, no. 12, 2022, doi: 10.3390/healthcare10122493.
- [5] D. C. Das dan M. S. Akter, "AI-Powered Multimodal Diagnostics in Modern Healthcare," *Asia Pacific J. Med. Innov.*, vol. 2, no. 3, 2025.
- [6] Z. Ahmad et al., "Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on pathology," *Diagn. Pathol.*, vol. 16, no. 1, 2021, doi: 10.1186/s13000-021-01085-4.
- [7] M. Caminati et al., "Biologics and global burden of asthma: A worldwide portrait and a call for action," *World Allergy Organ. J.*, vol. 14, no. 2, 2021, doi: 10.1016/j.waojou.2020.100502.
- [8] P. M. Pitrez, "The challenges of asthma care in low-and middle-income countries," *J. Bras. Pneumol.*, vol. 49, no. 3, 2023, doi: 10.36416/1806-3756/e20230215.
- [9] A. A. Faniyi et al., "Addressing the asthma crisis in Africa: challenges, strategies, and recommendations," *Egypt. J. Intern. Med.*, vol. 36, no. 1, 2024, doi: 10.1186/s43162-024-00340-6.
- [10] S. Sharma et al., "Vulnerability for Respiratory Infections in Asthma Patients: A Systematic Review," *Cureus*, vol. 14, no. 9, 2022, doi: 10.7759/cureus.28839.
- [11] R. D. Wagh et al., "Asthma: A Comprehensive Outlook," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 1, 2022.
- [12] V. E. Georgakopoulou et al., "Diagnostic tools in respiratory medicine," *Biomed. Reports*, vol. 23, no. 1, 2025, doi: 10.3892/br.2025.1990.
- [13] M. A. Almeshari et al., "The diagnosis of asthma: Can physiological tests of small airways function help?" *Chron. Respir. Dis.*, vol. 18, 2021, doi: 10.1177/14799731211053332.
- [14] Y. Lyu et al., "Non-invasive acoustic classification of adult asthma using an XGBoost model with vocal biomarkers," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-14645-1.
- [15] L. H. Lai et al., "The Use of Machine Learning Models with Optuna in Disease Prediction," *Electronics*, vol. 13, no. 23, 2024, doi: 10.3390/electronics13234775.
- [16] X. Wang et al., "A Joint Ensemble Framework for the Detection of Acute Exacerbations in COPD," *Res. Sq.*, 2023, doi: 10.21203/rs.3.rs-3712629/v1.
- [17] Z. Azmi et al., "Comparative Analysis of Hyperparameter Optimization Techniques on LightGBM for Asthma Prediction," *JITK*, vol. 11, no. 2, pp. 461-469, 2025, doi: 10.33480/jitk.v11i2.7369.
- [18] A. K. Muthevi dan V. M. Mutyam, "Using an Enhanced LightGBM Model to Predict Coronary Heart Disease: Performance Evaluation and Comparison," *Nanotechnol. Perceptions*, vol. 20, no. S14, 2024.
- [19] H. He dan E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
- [20] M. Branco et al., "A Survey of Predictive Modeling on Imbalanced Domains," *ACM Comput. Surv.*, vol. 49, no. 2, 2016, doi: 10.1145/2907070.

- [21] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nat. Mach. Intell.*, vol. 2, pp. 56-67, 2020, doi: 10.1038/s42256-019-0138-9.
- [22] X. Qi et al., "Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer," *Redox Biol.*, vol. 79, 2025, doi: 10.1016/j.redox.2024.103470.
- [23] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002.
- [24] G. E. A. P. A. Batista et al., "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20-29, 2004.
- [25] S. M. Lundberg dan S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [26] A. S. Listyoko et al., "Exploring the association between asthma and chronic comorbidities: impact on clinical outcomes," *Front. Med.*, vol. 11, 2024, doi: 10.3389/fmed.2024.1305638.
- [27] E. Y. A. Qaid dan I. Long, "Asthma unravelled: a comprehensive review of epidemiology, phenotypes, pathophysiology, and emerging therapies," *Egypt. J. Bronchol.*, vol. 19, no. 1, 2025, doi: 10.1186/s43168-025-00367-8.
- [28] J. Snoek et al., "Practical Bayesian Optimization of Machine Learning Algorithms," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.
- [29] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3146-3154, 2017.
- [30] R. Ranganathan et al., "Revisiting the optimal threshold for binary classification using machine learning," *J. Biomed. Inform.*, vol. 120, 2021, doi: 10.1016/j.jbi.2021.103846.
- [31] D. Toumpanakis dan O. S. Usmani, "Small airways in asthma: Pathophysiology, identification and management," *Chin. Med. J. Pulm. Crit. Care Med.*, vol. 1, no. 3, pp. 171-180, 2023, doi: 10.1097/CM9.0000000000002693.
- [32] A. S. Listyoko et al., "Association between obesity and asthma control: the role of BMI in respiratory outcomes," *Respir. Med.*, 2023.
- [33] R. Hurtado-Ruza et al., "Asthma, much more than a respiratory disease: Influence of depression and anxiety," *Rev. Assoc. Med. Bras.*, vol. 67, no. 4, pp. 571–576, 2021, doi: 10.1590/1806-9282.20201066.
- [34] J. A. Hanley dan B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.
- [35] P. K. Sahu dan T. Fatma, "Optimized Breast Cancer Classification Using PCA-LASSO Feature Selection and Ensemble Learning with Optuna Optimization," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3539746.
- [36] M. Khaldi et al., "Hyperparameter Optimization for Malicious URL Detection: Leveraging Optuna and Random Search in ML and DL Models," *Informatica*, vol. 49, no. 27, 2025, doi: 10.31449/inf.v49i27.9106.
- [37] H. Yang et al., "Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison," *IEEE Access*, vol. 11, pp. 23366-23380, 2023, doi: 10.1109/ACCESS.2023.3253885.
- [38] R. Sibundi et al., "A boosting ensemble learning based hybrid LightGBM and XGBoost model for predicting house prices," *Eng. Reports*, vol. 5, no. 4, 2023, doi: 10.1002/eng2.12599.
- [39] Q. Qian et al., "Prediction models for AKI in ICU: A comparative study," *Int. J. Gen. Med.*, vol. 14, pp. 623-632, 2021, doi: 10.2147/IJGM.S289671.
- [40] K.-T. Nguyen et al., "Research on the Influence of Genetic Algorithm Parameters on XGBoost in Load Forecasting," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 6, pp. 18849-18854, 2024, doi: 10.48084/etasr.8863.