Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.2, December 2024, pp. 296~303
e-ISSN: 2548-6861

Interpretable Machine Learning with SHAP and XGBoost for Lung Cancer Prediction Insights

Taufik Kurniawan [1]*, Laily Hermawanti [2], Achmad Nuruddin Safriandono [3]***
* Program Studi Sistem Komputer, Fakultas Teknik, Universitas Sultan Fatah, Demak Indonesia
** Program Studi Teknik Informatika, Fakultas Teknik, Universitas Sultan Fatah, Demak Indonesia
taufikkurniawan@unisfat.ac.id [1], lailyhermawanti18@gmail.com [2], udinozz@gmail.com [3]

Article Info

Article history:

Received 2024-08-27
Revised 2024-09-17
Accepted 2024-10-14

Keyword:

Balancing dataset,
Interpretable Machine Learning,
Lung cancer classification,
SHapley Additive exPlanations,
XGBoost.

ABSTRACT

Lung cancer remains one of the leading causes of death worldwide, and early detection through accurate and reliable methods is essential to improve patient prognosis. This study proposes a lung cancer classification model that integrates XGBoost with SHapley Additive exPlanations (SHAP) and Random Over Sampling (ROS) techniques to address the data imbalance problem. Using hyperparameter optimization through Optuna, the resulting model demonstrated superior performance, with an average accuracy of 96.84%, precision of 99.23%, recall of 94.51%, F1-score of 96.74%, specificity of 99.17%, and AUC of 96.84% in a 10-fold cross-validation evaluation. SHAP analysis provided significant interpretability, identifying key features such as gender, smoking habits, and physical signs of yellow fingers as the factors that most influence the model's predictions. The results of this study indicate that the proposed model is not only accurate, but also interpretable, making a significant contribution to supporting better clinical decision making in lung cancer diagnosis.

I. PENDAHULUAN

Kanker paru-paru tetap menjadi salah satu jenis kanker yang paling umum dan mematikan di seluruh dunia. Menurut Global Cancer Observatory, kanker paru-paru menyumbang 11,4% dari semua kasus kanker dan 18% dari semua kematian akibat kanker pada tahun 2020, menjadikannya penyebab utama kematian terkait kanker di dunia. Di Indonesia, misalnya, kanker paru-paru menyebabkan 30.843 kematian pada tahun 2020, dengan tingkat insiden yang lebih tinggi pada pria yang dipengaruhi oleh faktor-faktor seperti merokok[1]. Mengingat tingginya angka kematian dan seringkali keterlambatan diagnosis, ada kebutuhan yang mendesak untuk metode deteksi dini yang efektif untuk meningkatkan hasil pengobatan pasien[2]–[4].

Meskipun pembelajaran mesin telah menunjukkan potensi besar dalam diagnosis medis, kurangnya interpretabilitas dalam banyak model menjadi tantangan signifikan, terutama dalam lingkungan klinis di mana kepercayaan dan pemahaman sangat penting. Model pembelajaran mesin yang dapat dijelaskan, seperti yang memanfaatkan SHapley Additive exPlanations(SHAP), memberikan transparansi dengan memungkinkan tenaga medis untuk memahami proses pengambilan keputusan model[5]–[7]. Transparansi ini tidak hanya membantu dalam pengambilan keputusan klinis tetapi juga meningkatkan kepercayaan terhadap penggunaan teknologi AI dalam perawatan kesehatan.

Klasifikasi kanker paru-paru sangat penting untuk mendiagnosis dan membedakan antara kasus yang ganas dan non-ganas. Klasifikasi yang dini dan akurat dapat secara signifikan meningkatkan perencanaan pengobatan dan prognosis pasien. Berbagai algoritma pembelajaran mesin, termasuk Support Vector Machines (SVM) [1], Random Forest[8], K-nearest neighbors[9], Naive Bayes[10], Logistic Regression[11], Artificial Neural Network (ANN)[12], XGBoost[13]–[15], telah dieksplorasi karena potensinya dalam meningkatkan akurasi dan keandalan klasifikasi kanker paru-paru.

Di antara algoritma-algoritma ini, XGBoost menjadi pilihan yang menonjol karena kemampuannya untuk menangani dataset besar dan kompleks secara efisien, serta kemampuannya untuk melakukan otomatisasi dalam

penanganan fitur-fitur penting melalui pemangkasan pohon. XGBoost juga dikenal dengan kecepatan pelatihannya yang cepat dan kinerjanya yang kuat dalam berbagai tugas klasifikasi, terutama dalam kondisi data yang kompleks dan tidak seimbang. Selain itu, XGBoost memiliki fitur bawaan untuk menangani overfitting, seperti regularisasi, yang membuatnya sangat cocok untuk aplikasi medis di mana akurasi dan generalisasi adalah hal yang kritis[14]–[16].

Salah satu tantangan utama dalam klasifikasi kanker paru-paru adalah ketidakseimbangan dalam dataset, di mana jumlah kasus non-kanker seringkali jauh melebihi jumlah kasus kanker. Ketidakseimbangan ini dapat menyebabkan model yang bias yang terlalu terfokus pada kelas mayoritas, sehingga gagal mendeteksi kasus kelas minoritas dengan akurat, yang dalam hal ini adalah kasus kanker paru-paru. Mengatasi ketidakseimbangan ini melalui teknik seperti oversampling sangat penting untuk membangun model yang kuat[1], [17].

Random Oversampling (ROS) adalah salah satu teknik yang sederhana namun efektif untuk menangani masalah ini. Kelebihan utama dari ROS adalah kemampuannya untuk meningkatkan jumlah sampel pada kelas minoritas tanpa kehilangan informasi yang ada, sehingga memungkinkan model untuk belajar lebih baik tentang pola-pola pada kelas minoritas[18], [19]. Selain itu, teknik ini tidak mengubah data asli, yang berarti tidak ada informasi yang hilang atau terdistorsi. Meskipun metode ini dapat meningkatkan risiko overfitting, terutama dalam dataset yang sangat kecil, dalam konteks penggunaan algoritma seperti XGBoost, overfitting dapat dikendalikan melalui mekanisme regularisasi yang dimiliki oleh XGBoost, sehingga tetap menghasilkan model yang robust dan mampu memberikan hasil prediksi yang lebih akurat[13]–[15].

Metode yang diusulkan dalam penelitian ini bertujuan untuk meningkatkan akurasi dan interpretabilitas klasifikasi kanker paru-paru dengan mengintegrasikan SHAP dan XGBoost, serta mengatasi masalah ketidakseimbangan data menggunakan ROS. Pendekatan ini tidak hanya berusaha untuk meningkatkan kinerja model pada dataset yang tidak seimbang, tetapi juga memastikan bahwa proses pengambilan keputusan transparan dan dapat dipahami, sehingga memberikan wawasan yang berharga bagi klinisi. Penggunaan SHAP untuk analisis pentingnya fitur membantu dalam memahami fitur mana yang paling signifikan berkontribusi pada prediksi model, menjadikan hasil lebih dapat diandalkan dalam konteks klinis. Selain itu, dengan memanfaatkan ROS, metode ini meningkatkan jumlah sampel kelas minoritas tanpa mengorbankan informasi penting, dan bersama dengan mekanisme regularisasi XGBoost, risiko overfitting dapat dikendalikan, menghasilkan model yang robust dan akurat. Dengan demikian, penelitian ini tidak hanya berkontribusi pada peningkatan kualitas diagnosis kanker paru-paru, tetapi juga memperkuat penerapan pembelajaran mesin yang dapat dijelaskan dalam dunia medis, terutama dalam kasus-kasus dengan ketidakseimbangan data yang tinggi.

II. METODE

Seperti yang telah dijelaskan pada section sebelumnya penelitian ini bertujuan untuk meningkatkan akurasi dan interpretabilitas dalam klasifikasi kanker paru-paru, penelitian ini mengusulkan pendekatan yang mengintegrasikan XGBoost dengan SHAP untuk memastikan transparansi dalam pengambilan keputusan model. Selain itu, teknik ROS digunakan untuk menangani ketidakseimbangan data, sehingga model dapat lebih akurat dalam mendeteksi kasus kanker paru-paru. Tahapan dari metode yang diussulkan dijelaskan seperti dibawah ini:

A. Pengumpulan data

Dataset diambil dari [20], Dataset yang digunakan dalam penelitian ini terdiri dari 309 entri dengan 16 fitur. Tabel 1 menyajikan penjelasan mengenai fitur-fitur yang ada dalam dataset, termasuk tipe data untuk setiap fitur.

TABEL I
DETIL FITUR DATASET *LUNG CANCER* YANG DIGUNAKAN

No	Fitur	Tipe Data	Deskripsi
1	Gender	Object	Jenis kelamin responden (M untuk laki-laki, F untuk perempuan)
2	Age	Integer	Usia responden dalam tahun
3	Smoking	Integer	Status merokok (1: Tidak Merokok, 2: Merokok)
4	Yellow_fingers	Integer	Jari menguning (1: Tidak, 2: Ya)
5	Anxiety	Integer	Kecemasan (1: Tidak, 2: Ya)
6	Peer_pressure	Integer	Tekanan dari teman sebaya (1: Tidak, 2: Ya)
7	Chronic disease	Integer	Penyakit kronis (1: Tidak, 2: Ya)
8	Fatigue	Integer	Kelelahan (1: Tidak, 2: Ya)
9	Allergy	Integer	Alergi (1: Tidak, 2: Ya)
10	Wheezing	Integer	Mengi (1: Tidak, 2: Ya)
11	Alcohol consuming	Integer	Konsumsi alkohol (1: Tidak, 2: Ya)
12	Coughing	Integer	Batuk (1: Tidak, 2: Ya)
13	Shortness of breath	Integer	Sesak napas (1: Tidak, 2: Ya)
14	Swallowing difficulty	Integer	Kesulitan menelan (1: Tidak, 2: Ya)
15	Chest pain	Integer	Nyeri dada (1: Tidak, 2: Ya)
16	Lung_cancer	Object	Diagnosis kanker paru-paru (YES: Positif, NO: Negatif)

Dataset ini memiliki dua kelas dengan ditribusi 270 entri untuk kelas positif kangker paru-paru dan 39 entri untuk kelas negatif. Jika menganalisis lebih jauh kelas positif sekitar 87,3% dan kelas negatif 12.6%. Sehingga dataset ini dapat dikatakan sebagai dataset dengan kelas tidak seimbang.

B. Prapengolahan Penghapusan Nilai Hilang dan Data Duplikat

Tujuan dari tahapan menghapus data dupliat atau memiliki nilai duplikat adalah untuk meningkatkan integritas dan kualitas dataset sehingga model pembelajaran mesin yang dilatih dapat memberikan hasil yang lebih valid, akurat, dan dapat diandalkan. Dengan memastikan data bebas dari nilai yang hilang dan duplikat, serta memahami distribusi awal

kelas, peneliti dapat membangun model yang lebih robust yang mampu memberikan prediksi yang lebih baik.

C. Prapengolahan Label Encoding

Proses Label Encoding adalah teknik yang digunakan untuk mengubah variabel kategorikal menjadi nilai numerik, yang diperlukan karena sebagian besar algoritma pembelajaran mesin hanya dapat bekerja dengan data numerik. Pada dataset kanker paru-paru yang sedang kita bahas, ada dua fitur kategorikal yang perlu di-*encode* agar model pembelajaran mesin dapat memprosesnya: *Gender* dan *Lung_Cancer*.

Pada fitur *Gender* terdiri dari nilai "M" dan "F", sedangkan *Lung_cancer* terdiri dari "YES" dan "NO". Pada Gender "M" dapat diubah menjadi 0 "F" dan dapat diubah menjadi 1. Sedangkan pada *Lung_cancer* "YES" dapat diubah menjadi 1 dan "NO" dapat diubah menjadi 0. Data numerik yang dihasilkan dari Label Encoding dapat langsung digunakan dalam proses pelatihan model tanpa memerlukan konversi tambahan. Ini sangat krusial ketika model yang digunakan adalah model berbasis pohon keputusan seperti XGBoost, di mana pengambilan keputusan bergantung pada nilai numerik fitur input.

D. Prapengolahan Oversampling

Random Over Sampling (ROS) digunakan untuk mengatasi ketidakseimbangan kelas yang dapat menyebabkan model menjadi bias terhadap kelas yang lebih besar. ROS melakukan observasi dari kelas minoritas digandakan untuk menyeimbangkan distribusi kelas. Jika n_0 dan n_1 adalah jumlah observasi pada kelas 0 dan 1, dan $n_0 > n_1$. Maka setelah ROS, jumlah observasi menjadi $n'_1 = n_0$. Dimana n'_1 adalah jumlah observasi pada kelas 1 setelah oversampling.

E. Prapengolahan Normalisasi dan Reduksi Dimensi

Normalisasi dan reduksi dimensi dalam proses pembelajaran mesin bertujuan untuk meningkatkan kinerja model dengan membuat data lebih seragam dan sederhana. Normalisasi mengubah skala fitur agar setara, menghindari bias dan mempercepat konvergensi model[21]. Pada penelitian ini digunakan normalisasi *Standard Scaling* yang mengubah skala fitur dengan mengurangi rata-rata dan membaginya dengan standar deviasi. *Standard Scaling* dihitung dengan persamaan (1).

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

Dimana Z adalah data yang dinormalisasi, X adalah nilai asli, μ adalah rata-rata, dan σ standar deviasi.

Reduksi dimensi, melalui *Principal Component Analysis* (PCA), mengurangi jumlah fitur tanpa kehilangan informasi penting, mencegah overfitting, dan mempercepat waktu pelatihan[21]. Kedua langkah ini memastikan bahwa model lebih efisien, akurat, dan mudah diinterpretasikan. PCA dihitung dengan persamaan (2).

$$X_{pca} = X \cdot W \tag{2}$$

Dimana X_{pca} adalah data setelah PCA, X adalah data asli, dan W adalah matriks komponen utama. Sebagai catatan pada penelitian ini digunakan reduksi dimensi dengan PCA menjadi 12 komponen.

F. Desain Model dan Optimasi Hyperparameter dengan Optuna

Model yang digunakan dalam penelitian ini adalah XGBoost, sebuah algoritma berbasis *gradient boosting* yang telah terbukti efektif dalam berbagai tugas klasifikasi, terutama dalam menangani dataset yang kompleks dan tidak seimbang. XGBoost bekerja dengan membangun sekumpulan pohon keputusan secara berurutan, di mana setiap pohon baru berusaha untuk mengoreksi kesalahan prediksi yang dibuat oleh pohon sebelumnya. Model ini terkenal karena kecepatan pelatihan yang tinggi dan kemampuan untuk menangani dataset besar, serta fitur-fitur seperti regularisasi yang membantu mencegah overfitting.

Untuk mencapai kinerja optimal, *hyperparameter* dalam XGBoost perlu diatur secara tepat. *Hyperparameter* adalah parameter yang dikonfigurasi sebelum pelatihan model, berbeda dengan parameter model yang dipelajari dari data. Contoh hyperparameter dalam XGBoost meliputi jumlah pohon (n_estimators), kedalaman maksimum pohon (max_depth), dan tingkat pembelajaran (learning_rate). Pengaturan hyperparameter yang tepat sangat penting karena secara langsung memengaruhi kemampuan model untuk belajar dari data dan menggeneralisasi pada data baru.

Optimasi *hyperparameter* dilakukan menggunakan Optuna, sebuah pustaka optimasi berbasis *Bayesian* yang menggunakan algoritma *Tree-structured Parzen Estimator (TPE)*. Metode ini memanfaatkan informasi dari evaluasi sebelumnya untuk mengarahkan pencarian ke area yang lebih menjanjikan dalam ruang *hyperparameter*. Dengan cara ini, Optuna dapat menemukan pengaturan *hyperparameter* yang optimal dengan lebih efisien dibandingkan metode pencarian *grid* atau acak[22]. Tahapan optimasi pada penelitian ini dilakukan sebagai berikut:

1. Fungsi tujuan untuk optimasi adalah memaksimalkan skor *Area Under the ROC Curve* (AUC), yang merupakan ukuran kinerja yang penting untuk masalah klasifikasi dengan data tidak seimbang. Skor AUC dihitung dengan persamaan (3).

$$\text{AUC} = \int_0^1 TPR(FPR)dFPR \tag{3}$$

   Dimana TPR adalah *True Positive Rate* dan FPR adalah *False Positive Rate*.
2. Hyperparameter yang dioptimalkan:
   - n_estimators: Jumlah pohon keputusan yang digunakan dalam model.

- learning_rate: Kecepatan pembelajaran, mengontrol ukuran langkah saat model beradaptasi dengan kesalahan.
- max_depth: Kedalaman maksimum setiap pohon, mengontrol kompleksitas model.
- subsample: Proporsi sampel yang digunakan untuk setiap pohon, membantu dalam mencegah overfitting.
- colsample_bytree: Proporsi fitur yang digunakan untuk membangun setiap pohon, juga membantu dalam pencegahan overfitting.

3. Konfigurasi Optimasi: Optuna menjalankan beberapa percobaan dengan kombinasi *hyperparameter* yang berbeda, di mana setiap percobaan dinilai berdasarkan performa model pada set validasi *menggunakan cross-validation*. Pengaturan terbaik dipilih berdasarkan nilai AUC tertinggi yang dicapai di antara semua percobaan.
4. Model kemudian dilatih ulang menggunakan kombinasi *hyperparameter* terbaik ini dan dievaluasi pada data validasi.

G. Pelatihan dan Validasi Model

Seperti yang telah dijelaskan sebelumnya digunakan model XGBoost, sebuah algoritma yang menggabungkan prediksi dari banyak pohon keputusan yang lemah untuk membentuk model yang kuat melalui pendekatan *gradient boosting*. Dalam setiap iterasi, pohon keputusan baru ditambahkan untuk memperbaiki kesalahan prediksi dari model sebelumnya. Secara matematis, proses pembaruan model ini dapat dinyatakan dengan persamaan (4).

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{4}$$

Dimana $F_m(x)$ adalah model pada iterasi ke-m, $F_{m-1}(x)$ adalah model pada iterasi sebelumnya, $h_m(x)$ adalah pohon keputusan baru yang ditambahkan pada iterasi ke-m, dan γ_m adalah bobot pembelajaran yang mengontrol kontribusi dari pohon keputusan baru.

XGBoost dilatih menggunakan data yang telah dilakukan prapengolahn, termasuk normalisasi dan reduksi dimensi, untuk memastikan bahwa setiap fitur memiliki skala yang setara dan hanya fitur yang paling signifikan yang digunakan.

Untuk memastikan bahwa model yang dilatih dapat menggeneralisasi dengan baik ke data baru yang tidak terlihat selama pelatihan, kami menggunakan teknik *Stratified K-Fold Cross-Validation*. Teknik ini membagi dataset menjadi beberapa *fold* yang sama besar. Setiap *fold* digunakan sekali sebagai data uji, sementara *fold* lainnya digunakan sebagai data latih. Proses ini diulang sebanyak *K* kali, sehingga setiap observasi dalam dataset digunakan sebagai data uji tepat satu kali. Ini penting karena dataset yang digunakan tidak seimbang, dengan mayoritas besar berada pada kelas positif kanker paru-paru. Langkah-langkah validasi dilakukan sebagai berikut.

1. Dalam penelitian ini digunakan $K = 10$ *fold* untuk memaksimalkan pemanfaatan data yang terbatas. Selain itu juga digunakan parameter *random_state*=42.
2. Stratifikasi memastikan bahwa proporsi kelas dalam setiap *fold* sama dengan proporsi kelas dalam keseluruhan dataset.
3. Pada setiap iterasi, model dilatih pada $K - 1$ *fold* dan diuji pada fold yang tersisa. Hasil dari semua iterasi kemudian dirata-rata untuk memberikan estimasi kinerja model yang lebih akurat dan stabil. Metode evaluasi yang digunakan adalah metrik AUC pada persamaan (3).
4. Hasil AUC dari setiap iterasi validasi dirata-rata untuk mendapatkan estimasi kinerja model secara keseluruhan dengan persamaan (5).

$$CV(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{AUC}_k \tag{5}$$

Dimana $CV(\theta)$ adalah rata-rata AUC dari semua *fold*.

H. Evaluasi Kinerja Model

Metrik seperti *accuracy*, *precision*, *recall*, F1-*score*, dan *specificity* memberikan informasi tentang kinerja model. Matriks kebingungan memberikan pandangan detail tentang performa prediksi, sementara kurva ROC dan AUC mengukur trade-off antara TPR dan FPR pada berbagai threshold. Matrik-matrik tersebut dapat dihitung dengan persamaan (6)-(10)[23].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$f1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{9}$$

$$specificity = \frac{TN}{TN + FP} \tag{10}$$

I. Analisis Interpretabilitas Model dengan SHAP

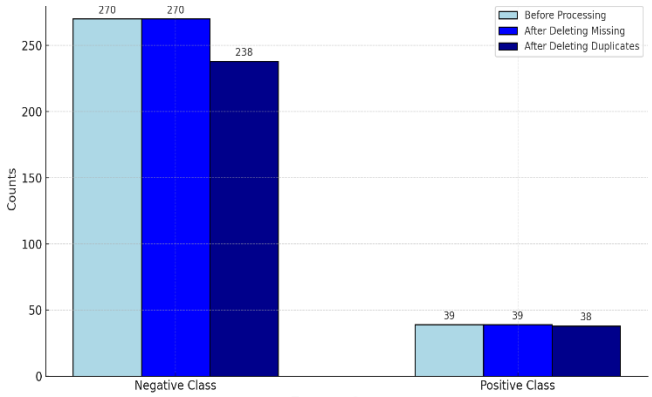
SHAP adalah metode untuk menjelaskan output dari model pembelajaran mesin. SHAP value untuk fitur i adalah kontribusi marjinal fitur tersebut terhadap prediksi dibandingkan dengan *baseline*. SHAP dapat dihitung dengan persamaan (11).

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{11}$$

Dimana ϕ_i adalah nilai SHAP untuk fitur i, S adalah subset dari semua fitur, dan $f(S)$ adalah fungsi prediktor.

III. HASIL DAN PEMBAHASAN

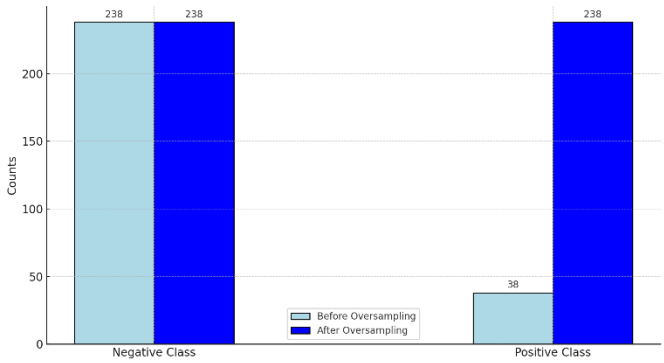
Pada tahap pertama setelah dataset dibaca dilakukan prapengolahan. Grafik pada Gambar 1 menunjukkan perubahan distribusi kelas dalam dataset pada tiga tahapan utama pengolahan data: sebelum pemrosesan, setelah penghapusan nilai yang hilang, dan setelah penghapusan data duplikat. Tidak ada perubahan dalam distribusi kelas setelah penghapusan nilai yang hilang, yang menunjukkan bahwa semua entri dalam dataset lengkap dalam variabel target. Setelah penghapusan data duplikat ada perubaha jumlah yaitu pada kelas negatif menurun dari 270 menjadi 238, sementara kelas positif dari 39 menjadi 38. Ini mengindikasikan bahwa beberapa entri yang identik telah dihapus, yang penting untuk memastikan bahwa model tidak dilatih pada data yang berlebihan atau tidak representatif.



Gambar 1. Plot perubahan dataset sebelum prapengolahan, setelah penghapusan nilai yang hilang, dan setelah penghapusan data duplikat.

Tahap selanjutnya dilaukan label *encoding* dan *oversampling* menggunakan ROS. Hasil proses ROS disajikan pada Gambar 2.

Proses *trial* dalam optimasi hyperparameter menggunakan Optuna bertujuan untuk menemukan kombinasi hyperparameter yang memberikan kinerja terbaik untuk model XGBoost. Proses trial dilakukan sebanyak 50 kali. Tabel 2 menyajikan ruang pencarian dan hasil *hyperparameter* terbaik.



Gambar 2. Plot perubahan dataset sebelum dan sesudan ROS.

TABEL II
RUANG PENCARIAN HYPERPARAMETER DAN KONFIGURASI TERBAIK BERDASARKAN OPTUNA

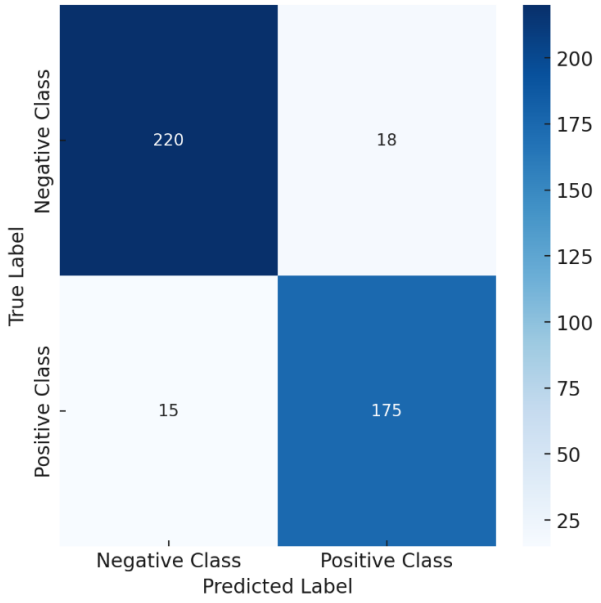
Hyperparameter	Ruang Pencarian	Konfiguarasi Terbaik
n_estimators	[100, 1000] (integer)	240
learning_rate	[0.01, 0.3] (float)	0.026844247528777843
max_depth	[3, 10] (integer)	9
subsample	[0.6, 1.0] (float)	0.8404460046972835
colsample_bytree	[0.6, 1.0] (float)	0.8832290311184181

Berdasarkan konfigurasi *hyperparameter* diatas, model XGBoost menghasilkan performa 10 *fold* yang disajikan pada Tabel 3. Grafik *confusion matrix* dan *receiver operating characteristic* (ROC), masing-masing disajikan pada Gambar 3 dan 4.

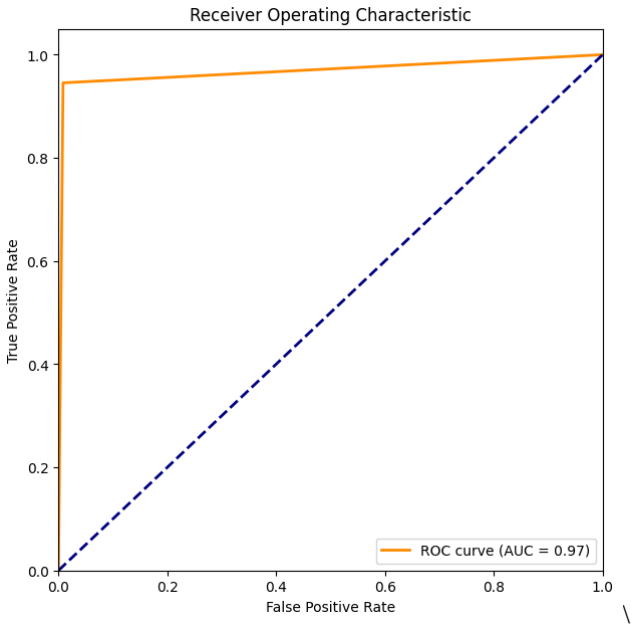
TABEL III
HASIL EVALUASI MODEL XGBOOST DENGAN KONFIGURASI TERBAIK OPTIMASI OPTUNA

Fold	Acc	Pre	Recall	F1	Spe	AUC
1	0,9583	1,0000	0,9167	0,9565	1,0000	0,9583
2	0,9792	1,0000	0,9583	0,9787	1,0000	0,9792
3	0,9583	0,9231	1,0000	0,9600	0,9167	0,9583
4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	0,9792	1,0000	0,9583	0,9787	1,0000	0,9792
6	0,9583	1,0000	0,9167	0,9565	1,0000	0,9583
7	0,9574	1,0000	0,9167	0,9565	1,0000	0,9583
8	0,9787	1,0000	0,9583	0,9787	1,0000	0,9792
9	0,9787	1,0000	0,9565	0,9778	1,0000	0,9783
10	0,9362	1,0000	0.8696	0,9302	1,0000	0,9348
avg	0,9684	0,9923	0,9451	0,9674	0,9917	0,9684

Berdasarkan hasil evaluasi kinerja model pada 10 fold cross-validation, model XGBoost yang dioptimasi menunjukkan performa yang sangat baik dalam klasifikasi kanker paru-paru. Akurasi rata-rata yang diperoleh adalah 96.84%, yang mengindikasikan bahwa model mampu memprediksi dengan benar sebagian besar data pada setiap fold. *Precision* yang mencapai 99.23% menunjukkan bahwa model memiliki tingkat kesalahan yang sangat rendah dalam memprediksi kasus positif, dengan sangat sedikit prediksi positif palsu. *Recall* rata-rata 94.51% menunjukkan bahwa model cukup sensitif dalam mendeteksi kasus positif, meskipun ada beberapa kasus positif yang tidak terdeteksi.



Gambar 3. Plot *confusion matrix* rata-rata dari 10 *fold*.



Gambar 4. Plot ROC rata-rata dari 10 *fold*.

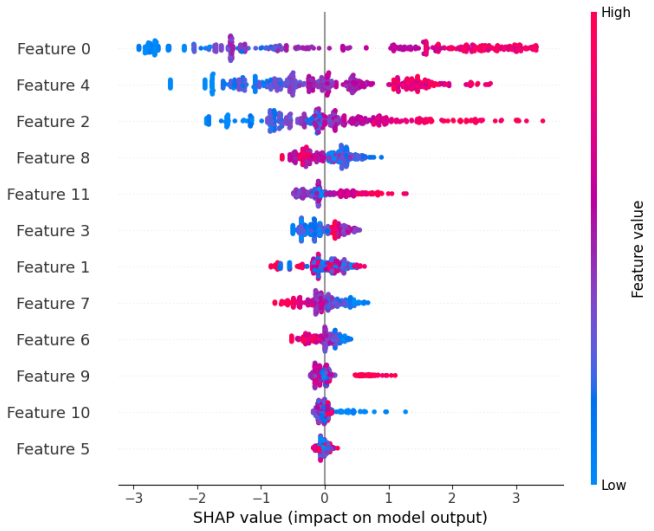
F1-Score rata-rata sebesar 96.74% menggabungkan keakuratan dan sensitivitas model, memberikan indikasi bahwa model seimbang dalam hal precision dan recall. Specificity rata-rata 99.17% menegaskan bahwa model sangat efektif dalam mengenali kasus negatif, dengan hampir semua kasus negatif diklasifikasikan dengan benar. AUC rata-rata sebesar 96.84% menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif.

Secara keseluruhan, hasil ini menunjukkan bahwa model XGBoost yang dioptimasi dengan teknik *cross-validation* ini sangat cocok untuk tugas klasifikasi kanker paru-paru, dengan keseimbangan yang baik antara precision, recall, dan akurasi, serta kemampuan yang kuat untuk menggeneralisasi performa pada data baru. Selanjutnya hasil metode yang diusulkan juga dilakukan komparasi dengan beberapa literatur terkait yang menggunakan dataset yang sama dimana disajikan pada Tabel 4.

TABEL IV
KOMPARASI DENGAN PENELITIAN TERKAIT PADA DATASET [20]

Alat Ukur	Referensi [12]	Referensi [2]	Referensi [4]	Metode Kami
Akurasi	0,93	0,93	95,4	0,9684
Presisi	0,91	-	95,4	0,9923
Recall	0,96	-	95,4	0,9451
F1-score	-	-	-	0,9674
Spesifisitas	0,90	-	-	0,9917
AUC	-	-	-	0,9684

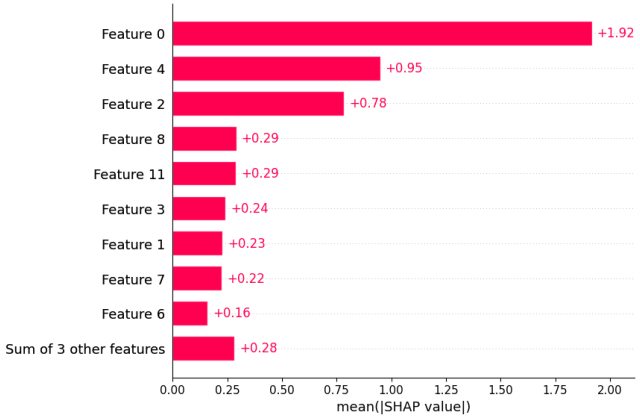
Metode yang diusulkan menunjukkan performa yang unggul dalam berbagai metrik dibandingkan dengan referensi lain. Dengan akurasi 96.84%, presisi 99.23%, dan spesifisitas 99.17%, metode ini terbukti lebih efektif dalam mengklasifikasikan data dengan benar dan mengurangi kesalahan positif palsu. Nilai recall dan F1-score yang tinggi juga mengindikasikan bahwa model ini seimbang dan andal dalam deteksi kasus positif dan negatif. Secara keseluruhan, metode ini menunjukkan kemampuan yang superior dalam tugas klasifikasi dibandingkan dengan penelitian sebelumnya.



Gambar 5. Pengaruh fitur-fitur utama terhadap prediksi model XGBoost berdasarkan SHAP.

Berdasarkan Gambar 5, SHAP yang disajikan, kita dapat melihat pengaruh relatif dari setiap fitur terhadap prediksi model XGBoost dalam klasifikasi kanker paru-paru. Warna pada grafik menunjukkan nilai fitur (biru rendah, merah tinggi), sementara sumbu horizontal menunjukkan nilai SHAP, yang mengukur dampak setiap fitur terhadap output model. Fitur dengan nilai SHAP yang lebih tinggi memiliki pengaruh lebih besar terhadap prediksi, baik dalam

meningkatkan atau menurunkan kemungkinan prediksi kanker paru-paru.



Gambar 6. Urutan pentingnya fitur dalam model prediksi berdasarkan SHAP.

Dari grafik tersebut, terlihat bahwa beberapa fitur seperti fitur 0 (*Gender*), fitur 4(*Yellow_finger*), dan fitur 2(*Smoking*) memiliki dampak yang signifikan pada prediksi model. Gambar 6 menunjukkan pentingnya fitur-fitur dalam model prediksi kanker paru-paru menggunakan nilai SHAP. Fitur 0 (Gender) memiliki dampak terbesar terhadap prediksi model, diikuti oleh Fitur 4 (Yellow_fingers) dan Fitur 2 (Smoking). Nilai SHAP yang lebih tinggi menunjukkan bahwa fitur tersebut lebih berpengaruh dalam menentukan hasil prediksi. Ini mengindikasikan bahwa gender, kebiasaan merokok, dan tanda-tanda fisik seperti jari kuning merupakan faktor utama yang dipertimbangkan model dalam mendiagnosis kanker paru-paru.

Interpretasi dari grafik ini penting bagi klinisi karena memungkinkan mereka untuk memahami bagaimana model membuat keputusan, dengan memberikan wawasan tentang fitur mana yang paling berkontribusi pada diagnosis kanker paru-paru. Dengan memahami pengaruh fitur-fitur ini, klinisi dapat meningkatkan interpretasi hasil model dalam konteks klinis dan memastikan bahwa prediksi model konsisten dengan pengetahuan medis yang ada.

IV. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa integrasi XGBoost dengan SHAP dan teknik ROS menghasilkan model klasifikasi kanker paru-paru yang tidak hanya akurat tetapi juga dapat diinterpretasikan. Dengan menerapkan optimasi hyperparameter menggunakan Optuna, model yang dihasilkan memiliki performa superior dibandingkan dengan metode lain, seperti yang ditunjukkan oleh hasil evaluasi menggunakan metrik-metrik utama seperti akurasi, presisi, recall, F1-score, spesifisitas, dan AUC. Selain itu, analisis interpretabilitas model menggunakan SHAP memberikan wawasan yang penting tentang fitur-fitur yang paling berpengaruh dalam prediksi, yang dapat digunakan oleh klinisi untuk meningkatkan pemahaman dan kepercayaan terhadap hasil prediksi yang diberikan oleh model. Secara keseluruhan, penelitian ini memperkuat peran penting pembelajaran mesin yang dapat dijelaskan dalam diagnosis medis, terutama dalam menangani data dengan ketidakseimbangan kelas yang tinggi.

UCAPAN TERIMA KASIH

Peneliti berterima kasih kepada Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi karena terlah mendukung penelitian ini dengan dana hibah dengan nomor 108/E5/PG.02.00.PL/2024

DAFTAR PUSTAKA

[1] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, “Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine,” *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 396–406, Mar. 2024, doi: 10.62411/jcta.10106.

[2] R. Yanuar, S. Sa’adah, and P. E. Yunanto, “Implementation of Hyperparameters to the Ensemble Learning Method for Lung Cancer Classification,” *Build. Informatics, Technol. Sci.*, vol. 5, no. 2, pp. 498–508, Sep. 2023, doi: 10.47065/bits.v5i2.4096.

[3] Y. F. Zamzam, T. H. Saragih, R. Herteno, Muliadi, D. T. Nugrahadi, and P.-H. Huynh, “Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based,” *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 125–136, Mar. 2024, doi: 10.35882/jeeemi.v6i2.382.

[4] E. Dritsas and M. Trigka, “Lung Cancer Risk Prediction with Machine Learning Models,” *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 139, Nov. 2022, doi: 10.3390/bdcc6040139.

[5] T. R. Noviandy, G. M. Idroes, and I. Hardi, “An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP,” *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 84–95, Aug. 2024, doi: 10.62411/faith.2024-16.

[6] R. K. Pathan, I. J. Shorna, M. S. Hossain, M. U. Khandaker, H. I. Almohammed, and Z. Y. Hamd, “The efficacy of machine learning models in lung cancer risk prediction with explainability,” *PLoS One*, vol. 19, no. 6, p. e0305035, Jun. 2024, doi: 10.1371/journal.pone.0305035.

[7] S. T. Rikta, K. M. M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, and S. K. Dey, “XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer,” *J. Pathol. Inform.*, vol. 14, p. 100307, Jan. 2023, doi: 10.1016/j.jpi.2023.100307.

[8] M. I. Akazue, I. A. Debekeme, A. E. Edje, C. Asuai, and U. J. Osame, “Unmasking Fraudsters: Ensemble Features Selection to Enhance Random Forest Fraud Detection,” *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 201–211, Dec. 2023, doi: 10.33633/jcta.v1i2.9462.

[9] M. A. Araaf, K. Nugroho, and D. R. I. M. Setiadi, “Comprehensive Analysis and Classification of Skin Diseases based on Image Texture Features using K-Nearest Neighbors Algorithm,” *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 31–40, Sep. 2023, doi: 10.33633/jcta.v1i1.9185.

[10] A. Wibowo and H. Hariyanto, “Comparison of Naive Bayes Method with Support Vector Machine in Helpdesk Ticket Classification,” *J. Appl. Informatics Comput.*, vol. 7, no. 2, pp. 165–171, Nov. 2023, doi: 10.30871/jaic.v7i2.6376.

[11] A. N. Safriandono, D. R. I. M. Setiadi, A. Dahlan, F. Z. Rahmanti, I. S. Wibisono, and A. A. Ojugo, “Analyzing Quantum Feature Engineering and Balancing Strategies Effect on Liver Disease Classification,” *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 51–63, Jun. 2024, doi: 10.62411/faith.2024-12.

[12] E. Vieira, D. Ferreira, C. Neto, A. Abelha, and J. Machado, “Data Mining Approach to Classify Cases of Lung Cancer,” in *Trends and Applications in Information Systems and Technologies*, 2021, pp. 511–521. doi: 10.1007/978-3-030-72657-7_49.

[13] F. Omoruwou, A. A. Ojugo, and S. E. Ilodigwe, “Strategic Feature

Selection for Enhanced Scorch Prediction in Flexible Polyurethane Form Manufacturing," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 346–357, Feb. 2024, doi: 10.62411/jcta.9539.

[14] R. E. Ako *et al.*, "Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.

[15] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Iriananda, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, May 2024, doi: 10.62411/faith.2024-11.

[16] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.

[17] D. R. I. M. Setiadi, H. M. M. Islam, G. A. Trisnapradika, and W. Herowati, "Analyzing Preprocessing Impact on Machine Learning Classifiers for Cryotherapy and Immunotherapy Dataset," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 39–50, Jun. 2024, doi: 10.62411/faith.2024-2.

[18] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *J. Big Data*, vol. 11, no. 1, p. 7, Jan. 2024, doi: 10.1186/s40537-023-00857-7.

[19] T. Riston *et al.*, "Oversampling Methods for Handling Imbalance Data in Binary Classification," in *Computational Science and Its Applications – ICCSA 2023 Workshops*, 2023, pp. 3–23. doi: 10.1007/978-3-031-37108-0_1.

[20] M. A. Bhat, "Lung Cancer Classification Dataset." Nov. 05, 2023. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

[21] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, and A. J. Tallón-Ballesteros, "The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis," in *18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023)*, 2023, pp. 344–353. doi: 10.1007/978-3-031-42536-3_33.

[22] S. Watanabe, "Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance," *arXiv*. Apr. 21, 2023. [Online]. Available: http://arxiv.org/abs/2304.11127

[23] T. R. Noviandy, K. Nisa, G. M. Idroes, I. Hardi, and N. R. Sasmita, "Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 358–367, Mar. 2024, doi: 10.62411/jcta.10129.