

ANALISIS PERFORMA ALGORITMA XGBOOST PADA KLASIFIKASI KANKER PARU-PARU

Maria Ere¹⁾, Danang Aditya Nugraha²⁾, Heri Santoso³⁾

Universitas PGRI Kanjuruhan Malang^{1,2,3)}

Email : eremersy@gmail.com

Abstrak

Penelitian ini bertujuan untuk menganalisis performa algoritma Extreme Gradient Boosting (XGBoost) dalam klasifikasi kanker paru-paru. Pengujian ini menggunakan dataset yang diperoleh dari kaggle dengan jumlah 220.632 data dan 23 atribut. Metode penelitian ini meliputi eksplorasi, preprocessing data seperti pembersihan, seleksi, dan transformasi data, serta pembagian data training dan testing sebanyak lima kali (70%:30%, 75%:25%, 80%:20%, 85%:15%, 90%:10%). Metric evaluasi model seperti akurasi, presisi, recall, dan f1-score digunakan melalui confusion matrix. Berdasarkan hasil pengujian model XGBoost dengan pengaturan parameter dapat menunjukkan performa yang baik meskipun proporsi data training dan testing diubah dengan akurasi tertinggi sebesar 95,76%, presisi sebesar 96%, recall sebesar 100% dan f1-score sebesar 98%.

Kata Kunci : Kanker Paru-paru; XGBoost; klasifikasi; machine learning; Confusion Matrix

Abstract

This study aims to analyse the performance of the Extreme Gradient Boosting (XGBoost) algorithm in lung cancer classification. This test used a dataset obtained from Kaggle, comprising 220,632 data points and 23 attributes. The research methods included exploration, data preprocessing such as cleaning, selection, and transformation, as well as dividing the data for training and testing five times (70%:30%, 75%:25%, 80%:20%, 85%:15%, 90%:10%). Model evaluation metrics such as accuracy, precision, recall, and F1-score were used in conjunction with the confusion matrix. Based on the results of XGBoost model testing with varying parameter settings, it demonstrates good performance even when the proportion of training and testing data is altered, achieving the highest accuracy of 95.76%, precision of 96%, recall of 100%, and F1-score of 98%.

Keywords : Lung Cancer, XGBoost, classification, machine learning, Confusion Matrix

1. PENDAHULUAN

Kanker paru-paru merupakan salah satu penyakit paling mematikan dimana sel-sel kanker berkembang pada organ paru-paru sehingga menyebabkan kematian yang diakibatkan oleh kanker di seluruh dunia. Menurut data World Health Organization (WHO), kanker paru-paru menyebabkan kematian sekitar 1,8 juta, sehingga menjadikannya penyebab kematian akibat kanker tertinggi secara global [1] dan terdapat 34.783 kasus baru dan Tingkat mortalitas sebanyak 30.843 kasus yang terjadi di Indonesia berdasarkan data GLOBOCAN 2020 dan diperkirakan akan terus bertambah setiap tahunnya [2].

Menurut Dr. Siswanto dari Rumah Sakit Akademik UGM, apa bila kanker paru-paru dapat terdeteksi pada stadium 1, maka tingkat kesembuhan akan mencapai 85%, tetapi jika

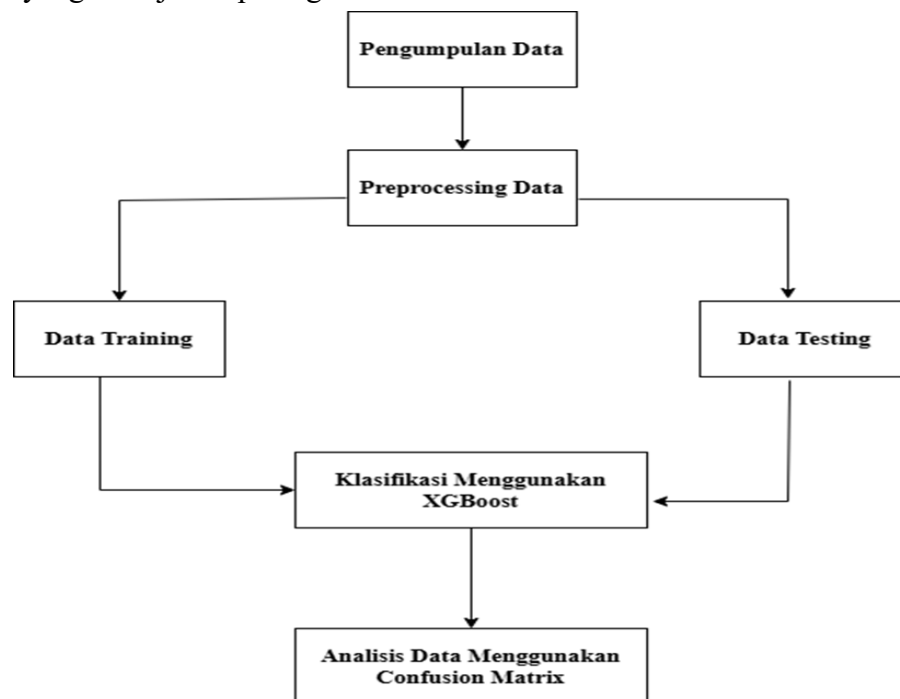
terdeteksi sudah pada stadium 4, maka tingkat kesembuhannya hanya 15%, sehingga perlunya deteksi dini dan diagnosis yang akurat agar meningkatkan penanganan yang lebih baik dan cepat untuk memperoleh peluang dalam kesembuhan dan mengurangi angka kematian yang diakibatkan oleh penyakit tersebut [3].

Dalam perkembangan kemajuan teknologi serta teknik analisis data sehingga mendorong pemanfaatan algoritma machine learning pada bidang kesehatan. Penelitian yang dilakukan oleh [4] dengan menggunakan algoritma *K-Nearest Neighbor* pada klasifikasi kanker paru-paru yang dapat menghasilkan akurasi sebesar 80.40%. Meskipun algoritma ini memberikan hasil yang cukup baik, namun memiliki keterbatasan dalam menangani data yang kompleks dan besar. Sehingga dibutuhkan metode lain seperti algoritma *Extreme Gradient Boosting* (XGBoost).

Algoritma Extreme Gradient Boosting (XGBoost) diperkenalkan pertama kali oleh Tianqi Chen pada tahun 2014 dan bekerja dengan cara menggabungkan pohon keputusan serta teknik gradient boosting untuk menciptakan model prediksi yang lebih kuat. Sehingga memiliki efisiensi, fleksibilitas, dan portabilitas tinggi untuk menyelesaikan masalah klasifikasi dan regres [5]. Penelitian ini bertujuan untuk mengklasifikasi kanker paru-paru menggunakan metode XGBoost, yang diharapkan hasilnya dapat berperan penting dalam ruang lingkup data medis.

2. METODE

Pada penelitian ini, menggunakan data yang bersumber dari kaggle tentang penyakit kanker paru-paru dan akan diolah menggunakan metode XGBoost. tahapan umum untuk penelitian ini yaitu seperti yang ditunjukkan pada gambar berikut ini:



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

pengumpulan data yang diambil dalam penelitian ini ialah data penyakit kanker paru-paru dengan jumlah sebanyak 220632 data yang memiliki 23 atribut. Berikut ini adalah deskripsi atribut yang digunakan dalam penelitian ini yang dapat dilihat pada tabel 2.1 berikut:

Tabel 2. 1 Deskripsi Atribut

No	Atribut	Keterangan
1	ID	Pengenal unik untuk setiap rekaman dalam kumpulan data
2	Negara	Negara tempat data dikumpulkan atau tempat tinggal individu
3	Ukuran populasi	Total ukuran populasi negara atau wilayah yang diteliti
4	Usia	Usia individu (dalam tahun) yang disertakan dalam data
5	Jenis kelamin	Jenis kelamin individu (misalnya, Pria, Wanita)
6	Perokok	Menunjukkan apakah individu tersebut seorang perokok (Ya/Tidak)
7	Tahun merokok	Jumlah tahun seseorang telah merokok
8	Rokok per hari	Jumlah rata-rata rokok yang dihisap seseorang setiap hari
9	Perokok pasif	Apakah individu tersebut terpapar asap rokok orang lain (Ya/Tidak)
10	Sejarah keluarga	Menunjukkan apakah individu tersebut memiliki riwayat keluarga kanker paru-paru (Ya/Tidak)
11	Diagnosis kanker paru-paru	Menunjukkan apakah Kanker Paru telah didiagnosis (Ya/Tidak)
12	Stadium kanker	Fitur pasca-diagnosis yang menunjukkan stadium 1-4 (NaN untuk semua yang didiagnosis negatif)
13	Tahun bertahan hidup	Tahun tindak lanjut pasca-diagnosis 1-10 (0 untuk semua yang didiagnosis negatif)
14	Jenis adenokarsinoma	jenis kanker yang berkembang di sel kelenjar yang menghasilkan lendir (Ya/Tidak)
15	Paparan polusi udara	objek kategoris dengan tiga tingkat: sedang, rendah, tinggi
16	Paparan pekerjaan	Terjadi paparan zat kimia, fisik, atau biologis berbahaya yang diterima pekerja selama menjalankan pekerjaannya (Ya/Tidak)
17	Polusi dalam ruangan	Terjadi polusi di dalam ruangan (Ya/Tidak)
18	Akses layanan kesehatan	kemampuan masyarakat untuk mendapatkan layanan kesehatan yang dibutuhkan (baik/buruk)
19	Deteksi dini	Melakukan upaya mengenali kondisi gangguan atau gejala dalam mengetahui adanya penyakit sejak dini (Ya/Tidak)
20	Maju atau berkembang	tingkat negara yang diteliti (Ya/Tidak)
21	Kematian tahunan akibat kanker paru-paru	Angka kematian pada setiap negara yang diteliti
22	Angka prevalensi kanker paru-paru	menunjukkan jumlah orang yang pernah didiagnosis menderita kanker paru-paru
23	Angka kematian	Angka kematian yang disebabkan oleh kanker paru-paru (NaN untuk semua yang didiagnosis negatif)

2.2 Ekplorasi Data

Sebelum melakukan preprocessing data sebaiknya terlebih dahulu melakukan eksplorasi data yang bertujuan untuk memahami struktur dari dataset yang diambil dari Kaggle. Dengan melakukan eksplorasi data dapat membantu dalam pengambilan keputusan saat melakukan preprocessing data.

2.3 Preprocessing Data

Preprocessing data dilakukan untuk mempersiapkan dataset sebelum digunakan dalam analisis sehingga menyesuaikan data dengan kebutuhan. preprocessing data yang dibutuhkan pada dataset ini yaitu sebagai berikut:

a. Seleksi data

Proses ini dilakukan untuk menghapus atribut-atribut yang tidak digunakan pada proses klasifikasi. Pada tahap ini atribut yang tidak berkaitan dengan diagnosis kanker paru-paru akan dihapus dari dataset.

b. Pembersihan data

Proses ini dilakukan untuk menghapus noise pada dataset. Pada langkah ini data-data yang tidak memiliki kelengkapan pada setiap atribut yang diperlukan maka akan dihapus dari dataset.

c. Transformasi data

proses ini dilakukan untuk mengubah data kategori kedalam bentuk numerik untuk dapat memudahkan dalam proses analisis. Teknik yang digunakan ialah Label Encoding untuk dataset ini.

2.4 Data Training dan Data Testing

Data training ditentukan sebagai data untuk percobaan model yang akan dilatih dan data testing ditentukan menjadi data yang digunakan untuk mengetahui kualitas model yang dilatih. Pada data penyakit kanker paru-paru ini penentuan data training dan data testing dilakukan secara random dengan lima kali percobaan dengan presentase berbeda yaitu 70% dan 30%, 75% dan 25%, 80% dan 20%, 85% dan 15% serta 90% dan 10%.

2.5 Analisis Data

Pada proses ini dilakukan analisis data dalam klasifikasi penyakit kanker paru-paru dengan metode XGBoost. Pada tahap ini dilakukan proses perhitungan untuk mendapatkan hasil akurasi yang lebih baik dengan melakukan pengujian menggunakan confusion matrix. Confusion matrix ialah salah satu metode evaluasi model yang cukup efektif dengan tujuan untuk memperoleh hasil akurasi yang lebih baik.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini data yang diperoleh terdapat sebanyak 220632 yang siap untuk dianalisis. Data tersebut mencakup 22 atribut diantaranya yaitu Jenis kelamin, Usia, Merokok, Jari kuning, Kecemasan, Tekanan teman, Penyakit kronis, Kelelahan, Alergi, Mengi, Konsumsi alkohol, Batuk, Sesak nafas, Kesulitan menelan dan Nyeri dada serta variabel target kanker paru-paru. Langkah pertama yang dilakukan adalah mengubah data mentah kedalam bentuk data CSV. Berikut gambar data mentah setelah diubah kebentuk CSV:

	ID	Country	Population_Size	Age	Gender	Smoker	Years_of_Smoking	Cigarettes_per_Day	Passive_Smoker	Family_History	...	Air_Pollution_Exposure
0	0	China	1400	80	Male	Yes	30	29	No	No	...	Low
1	1	Iran	84	53	Male	No	0	0	Yes	No	...	Low
2	2	Mexico	128	47	Male	Yes	12	6	Yes	No	...	Medium
3	3	Indonesia	273	39	Female	No	0	0	No	Yes	...	Low
4	4	South Africa	59	44	Female	No	0	0	Yes	No	...	Medium
...	...	...	...	...	...	...	...	...	...	...	...	...
220627	220627	Thailand	70	55	Male	Yes	4	10	No	Yes	...	Medium
220628	220628	South Africa	59	66	Male	No	0	0	No	No	...	High
220629	220629	Vietnam	97	46	Male	No	0	0	No	No	...	Medium
220630	220630	Ethiopia	120	36	Male	No	0	0	Yes	Yes	...	High
220631	220631	Germany	83	60	Female	No	0	0	No	No	...	Medium

220632 rows × 24 columns

Gambar 2. Data Hasil CSV

3.1 Preprocessing Data

Preprocessing data merupakan suatu proses persiapan data yang bertujuan untuk menghindari permasalahan dalam mengganggu hasil dari proses data mining. Dengan proses ini diharapkan dapat mempersiapkan data yang sesuai untuk kebutuhan analisis dan membuktikan kualitas data dengan baik. Berikut adalah gambar data setelah melalui proses preprocessing data:

	Age	Gender	Smoker	Years_of_Smoking	Cigarettes_per_Day	Passive_Smoker	Family_History	Lung_Cancer_Diagnosis	Adenocarcinoma_Type	Air_Pollution_Exposure	Occupational_Exposure	Indoor_Pollution	Early_Detection
0	80	1	1	30	29	0	0	0	1	1	1	0	0
1	53	1	0	0	0	1	0	0	1	1	1	0	0
2	47	1	1	12	6	1	0	0	1	2	0	0	1
3	39	0	0	0	0	0	1	0	1	1	0	0	0
4	44	0	0	0	0	1	0	0	1	2	1	0	0
--	--	--	--	--	--	--	--	--	--	--	--	--	--
220627	55	1	1	4	10	0	1	0	1	2	0	0	1
220628	66	1	0	0	0	0	0	0	0	0	0	0	0
220629	46	1	0	0	0	0	0	0	1	2	0	0	0
220630	36	1	0	0	0	1	1	0	1	0	0	0	0
220631	60	0	0	0	0	0	0	0	0	2	0	1	0

220632 rows x 13 columns

Gambar 3. Data Hasil Preprocessing Data

3.2 Data Training dan Data Testing

Data training akan digunakan untuk membentuk model data mining, sedangkan data testing akan digunakan dalam proses testing agar mengetahui performa model. Berikut tabel jumlah data setelah dibagi sesuai dengan presentase yang digunakan:

Tabel 3.1 Pembagian Data Training dan Data Testing

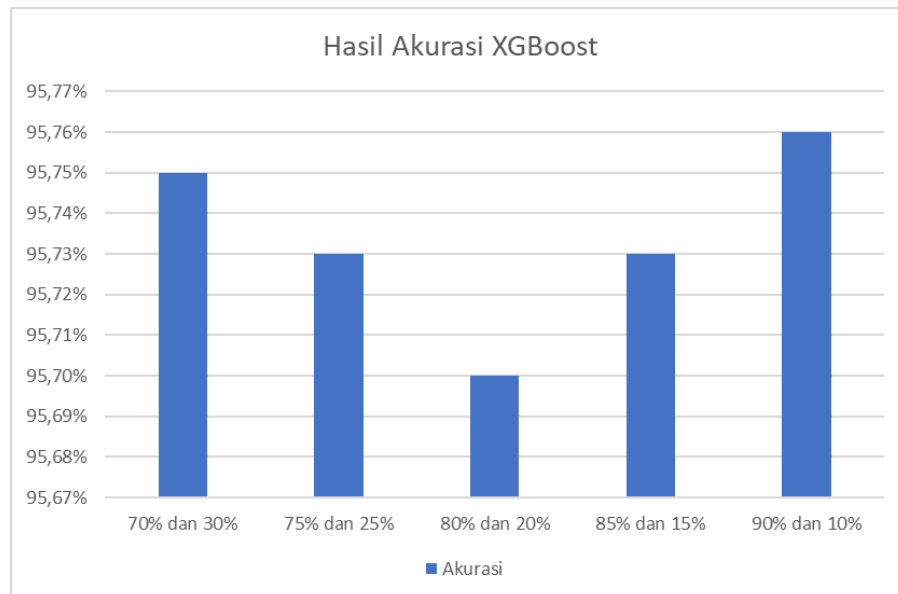
70% data training dan 30% data testing	75% data training dan 25% data testing	80% data training dan 20% data testing	85% data training dan 15% data testing	90% data training dan 10% data testing
Jumlah 70% = 154442	Jumlah 75% = 165474	Jumlah 80% = 176505	Jumlah 85% = 187537	Jumlah 90% = 198568
Jumlah 30% = 66190	Jumlah 25% = 55158	Jumlah 20% = 44127	Jumlah 15% = 33095	Jumlah 10% = 22064

3.3 Penerapan Metode XGBoost

Penerapan Algoritma XGBoost dalam penelitian digunakan untuk mengklasifikasi data kanker paru-paru dengan pengujian menggunakan Confusion Matrix. Pada pengujian algoritma XGBoost dengan melakukan percobaan sebanyak lima kali pada setiap masing-masing presentase, sehingga hanya mengambil hasil akurasi yang paling tinggi pada setiap percobaan. Hal ini juga terjadi pada hasil untuk metrik yang lain yaitu precision, recall dan F1-score untuk masing-masing presentase.

3.3.1 hasil akurasi

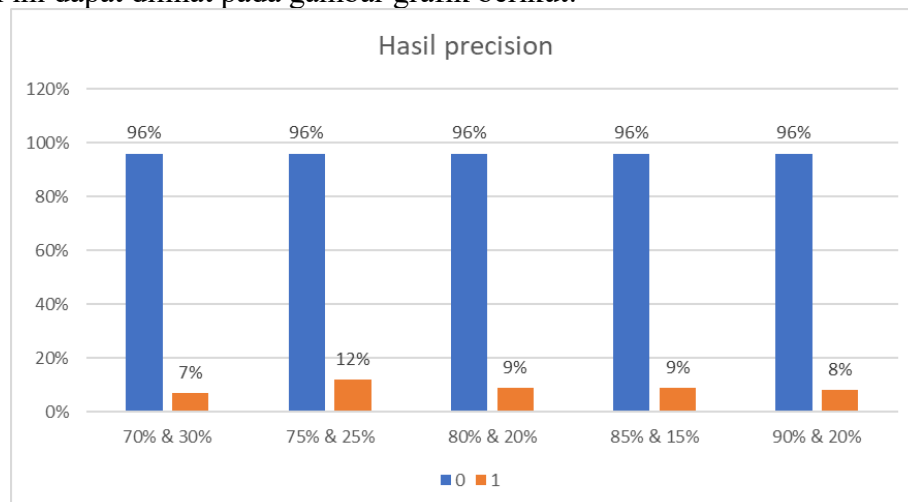
Akurasi yang didapatkan pada pengujian dengan metode XGBoost dengan masing-masing presentase data training dan data testing dapat memperoleh hasil yang tinggi, hasil ini dapat dilihat pada gambar grafik berikut:



Gambar 4. Grafik Hasil Akurasi

3.3.2 Hasil Precision

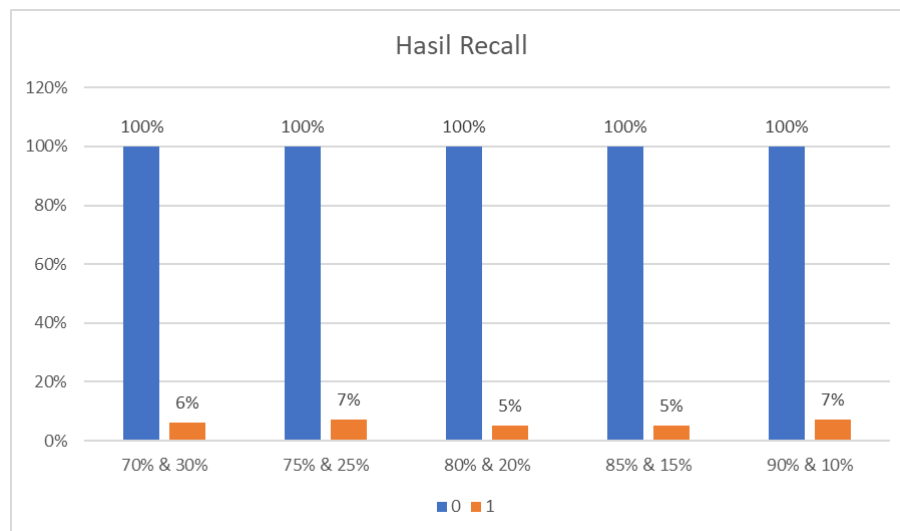
Hasil pengujian XGBoost pada metrik precision mendapatkan hasil yang sangat jauh berbeda antara kelas 0 (negatif) dan kelas 1 (positif). Precision pada kelas 0 menghasilkan kisaran yang sama yaitu 96%, sedangkan precision pada kelas 1 hanya mampu menghasilkan kisaran 7%-12%. Hasil ini dapat dilihat pada gambar grafik berikut:



Gambar 5. Grafik Hasil Precision

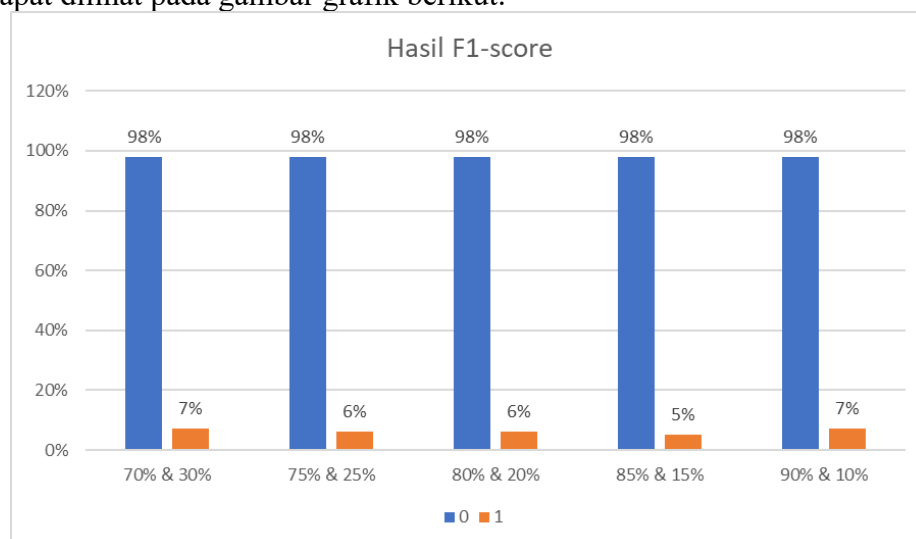
3.3.3 Hasil Recall

Hasil pengujian XGBoost pada metrik recall mendapatkan hasil yang sangat jauh berbeda antara kelas 0 (negatif) dan kelas 1 (positif). Recall pada kelas 0 menghasilkan kisaran yang sama yaitu 100%, sedangkan recall pada kelas 1 hanya mampu menghasilkan kisaran 5%-7%. Hasil ini dapat dilihat pada gambar grafik berikut:

**Gambar 6. Grafik Hasil Recall**

3.3.4 Hasil F1-score

Hasil pengujian XGBoost pada metrik f1-score mendapatkan hasil yang sangat jauh berbeda antara kelas 0 (negatif) dan kelas 1 (positif). F1-score pada kelas 0 menghasilkan kisaran yang sama yaitu 98%, sedangkan f1-score pada kelas 1 hanya mampu menghasilkan kisaran 5%-7%. Hasil ini dapat dilihat pada gambar grafik berikut:

**Gambar 7. Grafik Hasil Recall**

4. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian tersebut, maka diperoleh kesimpulan bahwa model XGBoost menunjukkan performa yang baik meskipun proporsi data training dan testing diubah. Hal ini dilihat dari hasil pengujian data training sebesar 90% dan data testing 10% menghasilkan akurasi tertinggi sebesar 95,76%, diikuti oleh pembagian 70% training dan 30% testing dengan akurasi 95,75%. Pengujian XGBoost untuk metrik lain seperti precision, recall dan f1-score yang tinggi masing-masing sekitar 96 persen hingga 100 persen untuk kelas negative (kelas 0), namun hanya menghasilkan 5% hingga 12% untuk kelas positif (kelas 1). Hasil ini menunjukkan bahwa model memiliki kecenderungan untuk ketidakseimbangan data sehingga menyebabkan perbedaan performa yang signifikan. Hal ini disebabkan oleh sampel kelas

negatif lebih banyak daripada sampel kelas positif, sehingga model tidak dapat menemukan pola di kelas minoritas, dan deteksi kasus positif menjadi kurang optimal.

Terdapat juga beberapa saran bagi penelitian selanjutnya dan pengembangan lebih lanjut adalah:

- a. Meskipun XGBoost telah terbukti menghasilkan akurasi yang cukup tinggi, penting untuk melakukan optimasi parameter yang lebih lanjut dengan menggunakan metode yang lebih sistematis seperti Grid Search, Random Search, atau Bayesian Optimization.
- b. Fokus pada hasil diatas, penelitian selanjutnya dapat mempertimbangkan penerapan teknik resampling dalam menangani ketidakseimbangan data seperti menggunakan metode oversampling (SMOTE atau ADASYN).

REFERENSI

- [1] WHO, “Kanker Paru-paru.” Accessed: Oct. 30, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/lung-cancer>
- [2] The International Agency for Research on Cancer, “GLOBOCAN 2020: Data Kanker Global Terbaru.” Accessed: Oct. 30, 2024. [Online]. Available: <https://www.uicc.org/news/globocan-2020-global-cancer-data>
- [3] N. Satria. Ardhi, “Deteksi Dini Kanker Paru-Paru.” Accessed: Nov. 11, 2024. [Online]. Available: <https://ugm.ac.id/id/berita/21946-deteksi-dini-kanker-paru-paru/>
- [4] T. Abdi Mangun, O. Nurdiawan, and A. Irma Purnamasari, “Lung Cancer Analysis Using K-Nearst Neighbor Algorithm,” 2023. [Online]. Available: https://ejournal.ubibanyuwangi.ac.id/index.php/jurnal_tinsika
- [5] A. H. Permana, F. R. Umbara, and F. Kasyidi, “Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting,” *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 1, p. 499, Jun. 2024.
- [6] M. Dava Maulana, A. Id Hadiana, and F. Rakhmat Umbara, “Algoritma Xgboost Untuk Klasifikasi Kualitas Air Minum,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 5, pp. 3251–3256, Jan. 2024, doi: 10.36040/jati.v7i5.7308.
- [7] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, “Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit,” *Journal of Mathematics: Theory and Applications*, pp. 21–26, Aug. 2022, doi: 10.31605/jomta.v4i1.1792.
- [8] M. A. Senubekti and L. A. Puspita Dewi, “Prinsip Klasifikasi dan Data Mining dengan Algoritma C4.5,” *Nuansa Informatika*, vol. 16, no. 2, pp. 87–93, Jul. 2022, doi: 10.25134/nuansa.v16i2.5834.
- [9] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, “Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang,” *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 15, no. 2, pp. 228–236, Dec. 2022, doi: 10.36456/jstat.vol15.no2.a5548.
- [10] T. Toharudin et al., “Boosting Algorithm to Handle Unbalanced Classification of PM2.5 Concentration Levels by Observing Meteorological Parameters in Jakarta-Indonesia Using AdaBoost, XGBoost, CatBoost, and LightGBM,” *IEEE Access*, vol. 11, pp. 35680–35696, 2023, doi: 10.1109/ACCESS.2023.3265019.
- [11] C. Zai, “Implementasi Data Mining Sebagai Pengolahan Data,” *Jurnal Portal Data*, vol. 2, no. 3, 2022, Accessed: Nov. 12, 2024. [Online]. Available: <http://portaldata.org/index.php/portaldata/article/view/107/109>