Efektivitas Algoritma Random Forest, XGBoost, dan Logistic Regression dalam Prediksi Penyakit Paru-paru

The Effectiveness of Random Forest, XGBoost, and Logistic Regression Algorithms in Predicting Lung Disease

Bernardus Septian Cahya Putra [1], Imam Tahyudin[2*], Bagus Adhi Kusuma[3], Khairunnisak Nur Isnaini[4]

[1,2,3,4]Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto

E-mail: [1]bernardusseptian30@gmail.com, [2*]imam.tahyudin@amikompurwokerto.ac.id, [3]bagus@amikompurwokerto.ac.id, [4]nisak@amikompurwokerto.ac.id

**Penulis korespondensi*

Abstrak

Penyakit paru-paru, seperti pneumonia dan kanker paru-paru, menjadi masalah kesehatan global dengan tingkat kematian tinggi, terutama dipengaruhi oleh polusi udara, infeksi, dan kebiasaan merokok. Pencegahan dan deteksi dini sangat penting dalam mengurangi dampaknya. Algoritma yang digunakan dalam penelitian ini meliputi Random Forest, XGBoost, dan Logistic Regression. Tujuannya yaitu untuk membandingkan performa tiga algoritma machine learning dalam mengklasifikasikan penyakit paru-paru menggunakan metrik evaluasi seperti, akurasi, presisi, recall, dan F1-score. Setelah hyperparameter tuning, XGBoost menunjukkan hasil terbaik dengan akurasi 94,44%, presisi 94,98%, recall 94,44%, dan F1-score 94,41%, menunjukkan keseimbangan optimal antara presisi dan recall. Random Forest juga memberikan hasil yang sebanding dengan XGBoost dengan akurasi dan presisi yang tinggi. Sementara itu, Logistic Regression menunjukkan keterbatasan dalam menangani data yang kompleks, dengan performa yang lebih rendah pada seluruh metrik evaluasi. Penelitian ini menunjukkan bahwa algoritma berbasis pohon keputusan seperti XGBoost dan Random Forest lebih unggul untuk klasifikasi penyakit paru-paru, menjadikannya metode yang lebih andal untuk mendukung deteksi dini penyakit ini.

Kata kunci: *Hyperparameter Tuning, Logistic Regression, Penyakit Paru-paru, Random Forest, XGBoost.*

Abstract

Lung diseases, such as pneumonia and lung cancer, are global health issues with high mortality rates, primarily influenced by air pollution, infections, and smoking habits. Prevention and early detection are essential in reducing the impact of these diseases. This study utilizes three machine learning algorithms, Random Forest, XGBoost, and Logistic Regression, with the aim of comparing their performance in classifying lung diseases using evaluation metrics such as accuracy, precision, recall, and F1-score. After hyperparameter tuning, XGBoost achieved the best results with an accuracy of 94.44%, precision of 94.98%, recall of 94.44%, and an F1-score of 94.41%, reflecting an optimal balance between precision and recall. Random Forest also showed results comparable to XGBoost, with high accuracy and precision. In contrast, Logistic Regression displayed limitations in handling complex data, with lower performance across all evaluation metrics. This study demonstrates that decision-tree-based algorithms like XGBoost and Random Forest are superior for lung disease classification, making them more reliable methods to support early disease detection.

Keywords: Hyperparameter Tuning, Logistic Regression, Lung Diseases, Random Forest, XGBoost.

1. PENDAHULUAN

Beberapa penyakit mematikan bisa terjadi pada paru-paru, salah satunya Pneumonia. Penyakit paru-paru, seperti pneumonia dan kanker paru-paru, merupakan masalah kesehatan serius yang dapat berujung pada kematian[1]. Kondisi ini sering disebabkan oleh paparan debu, asap, virus, atau bakteri yang menginfeksi saluran pernapasan. Kelompok usia yang rentan terinfeksi pun sangat beragam, mulai dari bayi hingga orang dewasa. Penyakit paru-paru sulit untuk disembuhkan dan berpotensi fatal, terutama di Indonesia, di mana kebiasaan merokok masih tinggi, bahkan di kalangan Masyarakat dengan berpenghasilan rendah, yang semakin memperburuk kesehatan paru-paru masyarakat[2]. Penyakit paru-paru, seperti kanker paru-paru dan pneumonia, menjadi masalah kesehatan global dengan angka kematian yang tinggi[3]. Deteksi dini dan upaya pencegahan sangat penting untuk menurunkan angka kematian. Saat ini, teknologi machine learning semakin banyak dimanfaatkan untuk mendukung diagnosis melalui analisis data dan citra medis, seperti X-ray dan CT scan[4], [5]. Algoritma seperti Random Forest, XGBoost, dan Logistic Regression telah terbukti efektif dalam mengklasifikasikan penyakit.

Algoritma Random Forest menggunakan sejumlah pohon keputusan untuk meningkatkan ketepatan dan stabilitas prediksi[1]. Sebagai teknik klasifikasi yang handal, Random Forest mampu menangani banyak variabel input tanpa mudah mengalami overfitting. Kelebihan terletak pada kemampuannya mengurangi korelasi antar pohon keputusan, yang menghasilkan prediksi lebih konsisten. Fleksibilitasnya memungkinkan algoritma ini digunakan dalam berbagai tugas, baik klasifikasi maupun regresi, serta dalam mengidentifikasi fitur-fitur penting dalam dataset pelatihan[6].

XGBoost, sebagai pengembangan dari Gradient Boosting, unggul dalam menangani data kompleks dan mengatasi masalah overfitting dengan regularisasi yang lebih baik[2]. Perbedaan antara Gradient Boosting dan XGBoost terletak pada proses penambahan "weak learner". Pada XGBoost, proses tersebut dilakukan secara multi-threaded, berbeda dengan Gradient Boosting yang dilakukan secara berurutan. Dalam situasi ini, penggunaan inti CPU dioptimalkan untuk meningkatkan kecepatan dan kinerja algoritma[7].

Sebagai alternatif yang lebih sederhana untuk klasifikasi biner, Regresi Logistik tetap menjadi pilihan utama. Algoritma ini dirancang untuk variabel respons biner dan dapat menerima prediktor berupa data kontinu, kategori, atau gabungan keduanya. Analisis ini tidak bergantung pada asumsi normalitas multivariat atau kesamaan matriks varian-kovarian, sehingga dapat diterapkan pada berbagai jenis skala data[8]. Regresi Logistik masih relevan dalam klasifikasi biner karena mampu memberikan interpretasi yang mudah dan efisien pada dataset dengan fitur linier[5].

Sejumlah penelitian terkait sebelumnya yang menjadi acuan dalam penelitian ini. Penelitian pertama tentang Klasifikasi Penyakit Paru-Paru Berdasarkan Peningkatan Kualitas Kontras dan EfficientNet Menggunakan Gambar X-Ray yang membahas masalah penting dalam mendiagnosis penyakit paru-paru, seperti tingginya angka kematian akibat keterlambatan deteksi, terbatasnya akses ke ahli radiologi yang berpengalaman, serta seringnya terjadi kesalahan diagnosis yang disebabkan oleh kualitas gambar medis yang kurang baik. Untuk mengatasi masalah kualitas gambar medis yang sering menjadi penyebab kesalahan diagnosis, teknik pra-pemrosesan seperti augmentasi, white balance, dan peningkatan kontras digunakan bersamaan dengan model EfficientNet untuk melakukan klasifikasi. Penelitian ini berfokus pada peningkatan akurasi dalam klasifikasi penyakit paru-paru, khususnya COVID-19, dengan memanfaatkan teknik peningkatan kontras CLAHE dan model pembelajaran mendalam EfficientNet. Hasil penelitian menunjukkan bahwa pendekatan ini memiliki potensi klinis yang signifikan, dengan F1-Score sebesar 0,97, recall 0,98, presisi 0,96, dan akurasi 0,97, yang menegaskan keefektifan model dalam klasifikasi multi-kelas[9].

Penelitian berikutnya tentang Klasifikasi Prediksi Penyakit Paru-Paru Normal dengan *Pneumonia* berdasarkan Citra Image *X-ray* dengan *Optimasi Adam Convolutional Neural Network (CNN)* yang membahas tentang beberapa tantangan besar dalam menentukan pneumonia dari gambar sinar-X paru-paru yang normal. Salah satunya adalah pentingnya diagnosis yang cepat untuk meningkatkan kesembuhan pasien. Studi ini bertujuan untuk mengembangkan model

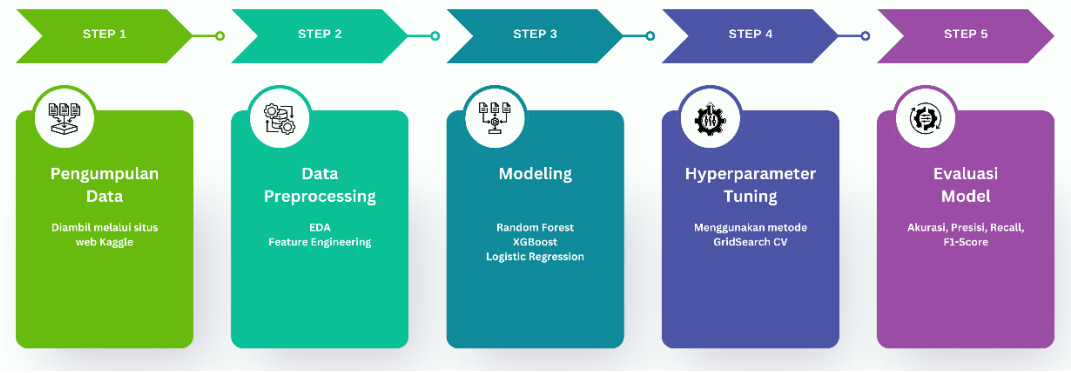
berbasis Convolutional Neural Networks (CNN) yang jelas membedakan antara paru-paru normal dan paru-paru yang terkena pneumonia. Dalam penelitian ini menggunakan algoritma CNN yang dioptimalkan menggunakan algoritma Adam dan dilatih dengan 624 gambar X-ray (234 normal, 390 pneumonia). Hasil penelitian menunjukkan bahwa sistem ini memiliki akurasi klasifikasi sebesar 92,15%, yang berarti bahwa pendekatan ini bisa meningkatkan diagnosis medis pada pasien ke depan secara signifikan[10].

Penelitian yang terakhir tentang Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques yang membahas tantangan-tantangan kritis untuk membedakan antara Covid-19 dan pneumonia pada gambar sinar-X dada, yang diperumit oleh penurunan kualitas gambar dan variabilitas dalam fitur yang diekstraksi. Tujuannya adalah untuk mengembangkan kerangka prognosa penyakit paru-paru yang akurat, dengan memberikan perhatian khusus kepada cara peningkatan akurasi pen-andikan dengan penghantaran sinyal kolomini secara lembut dan teknik normalisasi yang bubar. Beragam algoritma seperti SVM, ANN, KNN, dan pengklasifikasi ensemble telah diterapkan, dengan model F-RNN-LSTM yang berhasil mencapai akurasi 95%. Model ini secara signifikan lebih unggul dalam presisi, recall, dan F1-Score dibandingkan metode yang sudah ada, sambil juga mengurangi upaya komputasi sebesar 50%. Penelitian ini menekankan pentingnya teknik normalisasi yang kuat dalam meningkatkan kinerja klasifikasi dibandingkan dengan fitur mentah dan metode normalisasi lainnya[11].

Berdasarkan permasalahan di atas, penelitian ini bertujuan untuk membandingkan kinerja ketiga algoritma dalam mengklasifikasikan penyakit paru-paru dengan menggunakan metriks evaluasi seperti, akurasi, presisi, recall, dan F1-score. Dengan menggunakan dataset yang besar, diharapkan penelitian ini dapat memberikan analisis mendalam dan menentukan metode terbaik untuk deteksi dini penyakit paru-paru[1], [2].

2. METODE PENELITIAN

Beberapa tahapan penelitian mulai dari Dataset yang digunakan, Data Preprocessing, Modeling, Hyperparameter Tuning, dan yang terakhir Evaluasi yang bisa dilihat pada gambar 1.



Gambar 1 Flowchart Metode Penelitian

2.1 Pengumpulan data

Data yang digunakan pada penelitian ini adalah Dataset Predic Terkena Penyakit Paru-paru yang diambil dari web Kaggle dengan tautan berikut: https://www.kaggle.com/datasets/andot03bsrc/dataset-predic-terkena-penyakit-paruparu.

2.2 Data Preprocessing

Data Preprocessing adalah langkah penting dalam menyiapkan data untuk analisis yang valid, terutama dalam memprediksi penyakit paru-paru pada dataset yang besar. Tahapan ini meliputi pembersihan data untuk menghilangkan informasi yang tidak relevan, penanganan data yang hilang, transformasi data ke dalam format yang sesuai untuk pemrosesan, serta pembagian data guna evaluasi yang akurat. Teknik seperti imputasi untuk nilai yang hilang dan deteksi outlier

meningkatkan kualitas dataset, membantu mengurangi overfitting dan mengoptimalkan kinerja model seperti Random Forest, XGBoost, dan Logistic Regression dalam klasifikasi penyakit[2], [12].

2.3 Modeling

Dalam penelitian ini, dibandingkan kinerja algoritma Random Forest, XGBoost, dan Logistic Regression untuk memprediksi penyakit paru-paru. Prediksi ini didasarkan pada dataset yang mencakup faktor-faktor utama, seperti usia, jenis kelamin, bekerja, kebiasaan merokok, rumah tangga, aktivitas begadang, aktivitas olahraga, asuransi, penyakit bawaan[13].

(Revisi) – Berikut ada dua skenario evaluasi yang digunakan untuk membandingkan performa algoritma:

1. Evaluasi tanpa Hyperparameter Tuning

Pada skenario ini, ketiga algoritma diimplementasikan menggunakan pengaturan default tanpa melakukan tuning hyperparameter. Kinerja yang dihasilkan diukur menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score untuk memberikan baseline kinerja awal.

2. Evaluasi menggunakan Hyperparameter Tuning

Dalam skenario ini, hyperparameter tuning diterapkan pada setiap algoritma untuk meningkatkan kinerjanya. Metode Grid Search digunakan untuk pencarian kombinasi hyperparameter terbaik. Pada Random Forest, Hyperparameter yang dioptimalkan meliputi n_estimators, max_depth, min_samples_split, dan min_samples_leaf. Pada XGBoost, Parameter yang digunakan mencakup learning_rate, n_estimators, max_depth, subsample, dan colsample_bytree. Pada Logistic Regression, Hyperparameter seperti penalty, C, solver, dan max_iter diatur untuk memaksimalkan kinerja model. – (Revisi)

2.3.1 Random Forest

Random Forest adalah algoritma ensemble learning yang menggabungkan sejumlah decision tree untuk keperluan klasifikasi atau regresi, dengan tujuan meningkatkan akurasi dan mengurangi risiko overfitting[14]. Setiap pohon dalam model ini dibangun dari subset data dan atribut yang dipilih secara acak, sedangkan hasil akhir ditentukan melalui voting atau rata-rata dari semua pohon[15]. Algoritma ini dikenal memiliki akurasi tinggi, terutama pada data yang kompleks, dan sering kali lebih unggul dibandingkan Logistic Regression dalam studi medis[16]. Random Forest menggunakan metrik seperti entropy dan information gain untuk pemisahan yang optimal, sehingga mampu mengidentifikasi pola prediktif dengan tingkat presisi yang tinggi[15].

Dibawah ini adalah rumus untuk menentukan pohon keputusan:

Menghitung nilai entropy seperti yang tertera pada persamaan 1.

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

Keterangan:

S = Kumpulan dataset

n = Jumlah kelas yang ada

p_i = Probabilitas dari setiap kelas ke-i pada output S

Menghitung nilai information gain seperti yang tertera pada persamaan 2.

$$Information\ Gain(A) = Entropy(S) = \sum_{i=1}^{n} \frac{S_i}{S} x\ Entropy(S_i) \quad (2)$$

Keterangan:

A = Atribut

S = Kumpulan data

$|S_i|$ = Jumlah sampel untuk nilai ke-i

$|S|$ = Jumlah data yang besar

2.3.2 XGBoost

XGBoost, atau eXtreme Gradient Boosting, adalah pengembangan dari Gradient Boosting, dirancang untuk meningkatkan akurasi dan efisiensi dalam tugas klasifikasi dan regresi[17]. Dikembangkan oleh Tianqi Chen pada tahun 2014, algoritma ini mengurangi overfitting melalui fungsi regulasi yang efektif dan belajar secara bertahap untuk mengurangi

kesalahan[18]. XGBoost unggul dalam menangani variabel kategorikal dan kelas yang tidak seimbang, sehingga populer dalam bidang keuangan dan analisis risiko kesehatan berkat efisiensinya yang tinggi serta kemampuannya dalam penyesuaian parameter untuk kinerja yang optimal[19].

Berikut ini merupakan persamaan dari algoritma XGBoost[20]:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

Keterangan:

$L^{(t)}$ = Fungsi loss

n = Jumlah iterasi nilai dari model

y_i = Nilai target aktual

$\hat{y}_i^{(t-1)}$ = Prediksi kumulatif model hingga iterasi ke-t – 1

$f_t(x_i)$ = Fungsi tambahan pada iterasi ke-t

$\Omega(f_t)$ = Fungsi regularisasi yang mengukur kompleksitas model

2.3.3 Logistic Regression

Logistic Regression merupakan algoritma yang banyak digunakan dalam tugas klasifikasi biner, memprediksi probabilitas suatu kejadian dengan menggunakan fungsi sigmoid yang menghasilkan nilai antara 0 dan 1, ideal untuk klasifikasi "positif" dan "negatif"[21]. Algoritma ini cukup fleksibel untuk menangani variabel independen baik kontinu maupun diskrit, serta memungkinkan interpretasi langsung mengenai pengaruh variabel-variabel tersebut pada variabel dependen[22]. Dalam konteks prediksi penyakit, Logistic Regression populer karena ketepatan dan kecepatan prosesnya, meskipun rentan mengalami underfitting jika distribusi kelas pada data tidak seimbang[23].

Persamaan dalam regresi logistik berfungsi untuk memodelkan probabilitas dari hasil biner (1 atau 2). Berdasarkan data yang tersedia, model ini menerapkan fungsi logistik guna memprediksi kemungkinan terjadinya suatu hasil. Fungsi logistik tersebut dirumuskan sebagai[24]:

$$P(Y = 1) = \frac{1}{1+e-(b_0-b_1x_1+b_2x_2+\cdots+b_px_p} \quad (4)$$

Keterangan:

P = Probabilitas

x_1, x_2,..., x_p = Nilai prediktor

b_0, b_1,..., b_p= Intersepsi model

2.4 Hyperparameter Tuning

Hyperparameter tuning adalah proses menemukan kombinasi parameter yang optimal untuk meningkatkan kinerja model machine learning. Karena tidak dapat dipelajari langsung dari data, hyperparameter harus ditentukan sebelum pelatihan, baik secara manual maupun otomatis melalui metode seperti Grid Search, Random Search, atau Bayesian Optimization yang lebih efisien[25]. Hyperparameter bervariasi menurut algoritma yang digunakan. Pada XGBoost, hyperparameter seperti learning_rate, n_estimators, max_depth, subsample, dan colsample_bytree penting untuk pengaturan kecepatan pembelajaran dan pemilihan fitur. Dalam Random Forest, n_estimators, max_depth, min_samples_split, min_samples_leaf, dan bootstrap mengontrol jumlah dan kedalaman pohon, serta penggunaan data bootstrap. Pada Logistic Regression, hyperparameter seperti penalty, C, solver, dan max_iter mengatur regularisasi dan optimasi. Tuning optimal dari hyperparameter ini dapat secara signifikan meningkatkan performa model[26]. Pemilihan hyperparameter yang tepat dapat membantu model mencapai generalisasi yang baik dan meningkatkan akurasi pada data baru.

2.5 Evaluasi

Penelitian ini mengevaluasi model menggunakan metrik seperti akurasi, presisi, recall, dan F1 score. Metrik-metrik ini digunakan untuk membandingkan kinerja algoritma Random Forest, XGBoost, dan Logistic Regression bekerja ketika memprediksi penyakit paru-paru[27].

2.5.1 Akurasi

Akurasi digunakan untuk menilai kinerja suatu metode klasifikasi berdasarkan ketepatan hasil klasifikasinya. Semakin tinggi nilai akurasi yang diperoleh, maka metode tersebut dinilai semakin efektif. Cara menghitung akurasi dapat dilihat pada rumus berikut ini[28]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

2.5.2 Presisi

Presisi mencakup tingkat sensitivitas atau akurasi sistem dalam menilai informasi yang diberikan, sehingga dapat mengidentifikasi data positif atau negatif secara tepat. Berikut rumus untuk menghitung presisi[2]:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

2.5.3 Recall

Recall merupakan proporsi data positif yang terdeteksi secara akurat dibandingkan dengan seluruh data positif, baik yang benar maupun yang keliru dikategorikan sebagai negatif. Nilai recall dapat dihitung menggunakan persamaan sebagai berikut[28]:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

2.5.4 F1-Score

F1-Score adalah metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, khususnya saat data memiliki distribusi kelas yang tidak seimbang. Persamaan F1-Score bisa dilihat sebagai berikut[14]:

$$F1 - Score = 2x\frac{Precision\ x\ Recall}{Precision+Recall} \quad (8)$$

3 HASIL DAN PEMBAHASAN

Hasil dan pembahasan dalam penelitian ini disajikan sesuai dengan tahapan-tahapan metode penelitian yang telah dilaksanakan. Pada bagian ini, akan dijelaskan secara rinci hasil analisis data dan evaluasi model prediktif yang digunakan pada dataset.

3.1 Pengumpulan Data

Dataset ini diambil dari Kaggle yang berjumlah 30.000 baris dan 11 kolom, yang memberikan informasi komprehensif tentang faktor-faktor yang berkontribusi pada risiko terkena penyakit paru-paru.

	No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan	Hasil
0	1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak	Ya
1	2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada	Tidak
2	3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
3	4	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada	Tidak
4	5	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Ya
...	...	...	...	...	...	...	...	...	...	...	...
29995	29996	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
29996	29997	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
29997	29998	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
29998	29999	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Tidak
29999	30000	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya

30000 rows × 11 columns

Gambar 2 Pengumpulan data

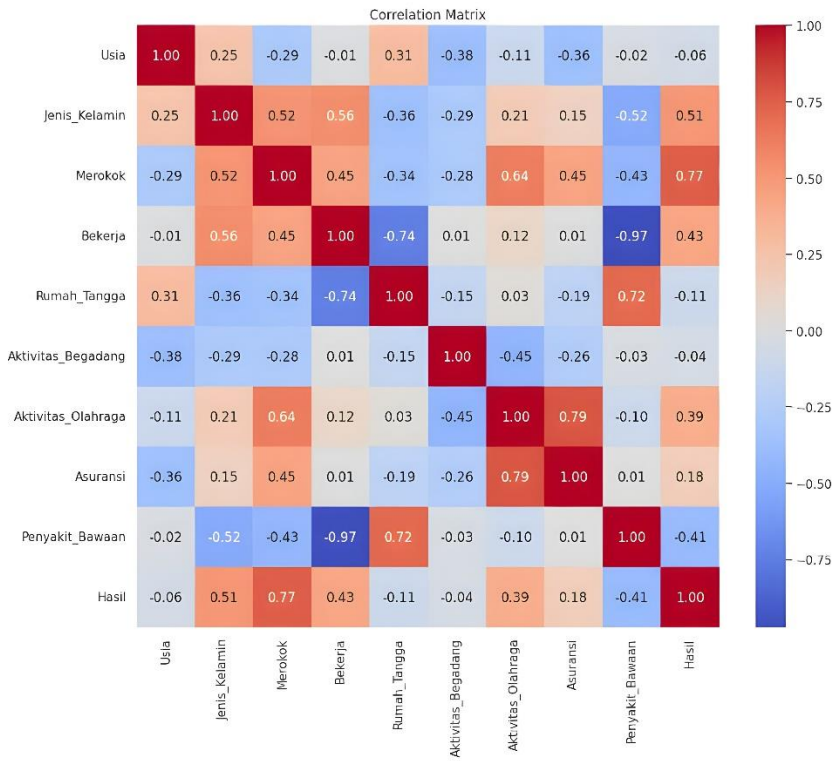
3.2 Feature Engineering

```
Data setelah Feature Engineering:
   Usia  Jenis_Kelamin  Merokok  Bekerja  Rumah_Tangga  Aktivitas_Begadang  \
0     1              0        1        0             1                   1
1     1              0        0        0             1                   1
2     0              0        0        0             1                   1
3     1              0        0        1             0                   0
4     0              1        1        1             0                   0

   Aktivitas_Olahraga  Asuransi  Penyakit_Bawaan
0                   1         0                1
1                   0         0                0
2                   0         0                1
3                   0         0                0
4                   1         1                0

Label pada variabel target (Hasil):
0    1
1    0
2    0
3    0
4    1
Name: Hasil, dtype: int64
```

Gambar 3. Feature Engineering

Gambar 3 menunjukkan hasil setelah proses feature engineering, data dikonversi menjadi format numerik untuk mempermudah analisis dan pemodelan. Fitur-fitur kategorikal seperti ‘Jenis_Kelamin’, ‘Merokok’, ‘Bekerja’, dan lainnya diubah menjadi nilai ‘0’ dan ‘1’ untuk menunjukkan keberadaan atau kondisi spesifik dari setiap atribut. Hal ini membantu meningkatkan kinerja algoritma machine learning karena model lebih mudah menangani data numerik, terutama model yang bergantung pada perhitungan matematis. Selain itu, variabel target ‘Hasil’ juga dilabeli secara numerik untuk memudahkan klasifikasi. Dengan hasil transformasi ini, dataset menjadi siap untuk digunakan dalam model prediksi, menjaga informasi dari setiap fitur sambil mengurangi kompleksitas yang mungkin muncul dari variabel kategorikal, sehingga diharapkan dapat menghasilkan prediksi yang lebih akurat dan efisien.

3.3 Correlation Matrix

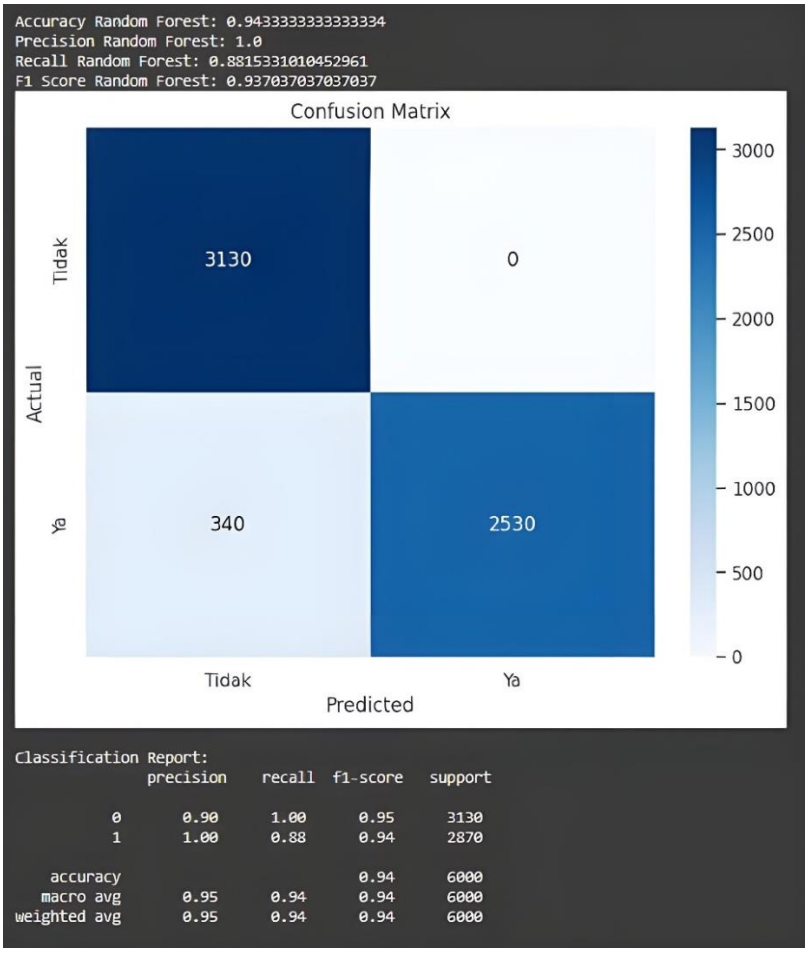


Gambar 4 Matriks Korelasi

Gambar 4 menampilkan matriks korelasi yang memvisualisasikan hubungan antar variabel dalam dataset, dengan nilai korelasi berkisar dari -1 hingga 1. Warna merah menunjukkan korelasi positif yang kuat, sedangkan biru menunjukkan korelasi negatif yang kuat. Korelasi positif berarti bahwa peningkatan pada satu variabel diikuti oleh peningkatan pada variabel lainnya, sedangkan korelasi negatif menunjukkan hubungan yang berlawanan. Sebagai contoh, variabel 'Merokok' memiliki korelasi positif tinggi dengan 'Hasil' (0,77), menunjukkan dampak signifikan merokok terhadap prediksi penyakit. Sementara itu, 'Penyakit_Bawaan' memiliki korelasi negatif dengan 'Bekerja' (-0,97). Matriks ini membantu mengidentifikasi fitur-fitur penting yang memengaruhi variabel target dan mengurangi risiko multikolinearitas yang dapat memengaruhi performa model.

3.4 Evaluasi model

3.4.1 Algoritma Random Forest

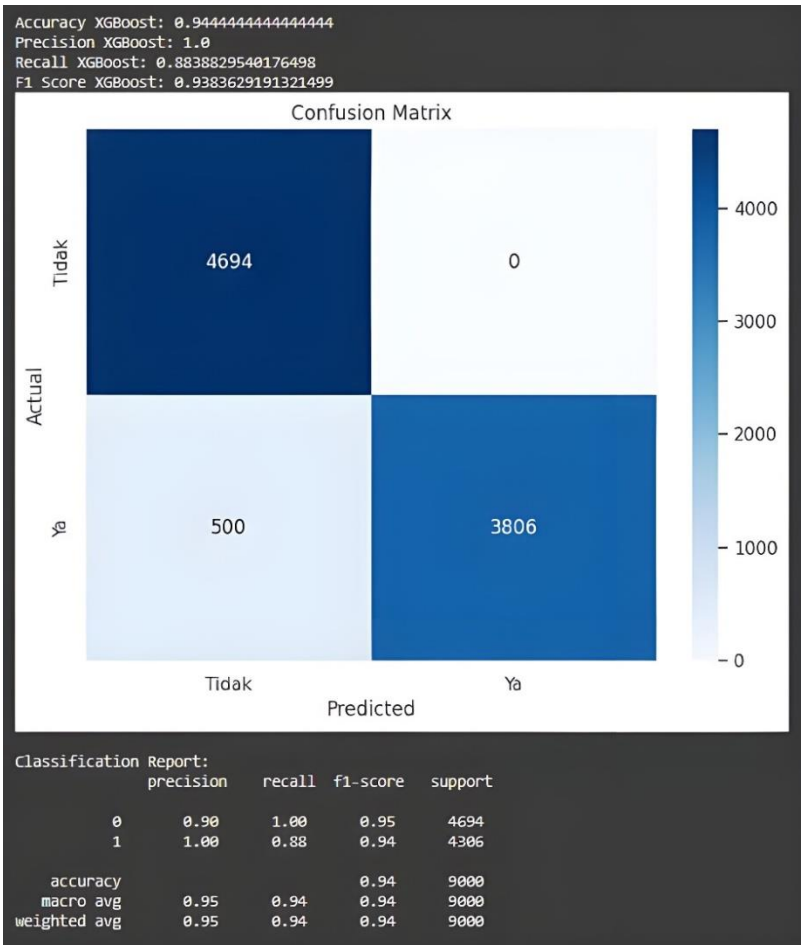


Gambar 5 Evaluasi Model Random Forest

Evaluasi model Random Forest pada gambar 5 menunjukkan performa yang cukup baik dalam melakukan klasifikasi, dengan akurasi sebesar 94,3%, presisi sebesar 100%, recall sebesar 88,15%, dan F1 Score sebesar 93,70%. Confusion matrix mengungkapkan bahwa model berhasil mengklasifikasikan 3.130 sampel kelas "Tidak" dan 2.530 sampel kelas "Ya" secara tepat. Tidak ditemukan kesalahan prediksi di mana kelas "Tidak" diprediksi sebagai "Ya" (False Positive), namun terdapat 340 sampel dari kelas "Ya" yang salah diklasifikasikan sebagai "Tidak" (False Negative). Berdasarkan classification report, nilai precision untuk kelas 0 sebesar 0,90, sementara untuk kelas 1 mencapai 1,00. Recall pada kelas 0 mencapai 1,00, yang berarti seluruh sampel kelas 0 diklasifikasikan dengan benar, sedangkan recall untuk kelas 1 sebesar 0,88, menunjukkan adanya beberapa sampel kelas 1 yang tidak teridentifikasi dengan tepat. Nilai F1-score

keseluruhan untuk kedua kelas adalah 0,94, menunjukkan performa model yang baik, meskipun masih ada sedikit ruang untuk perbaikan, terutama pada recall kelas 1.

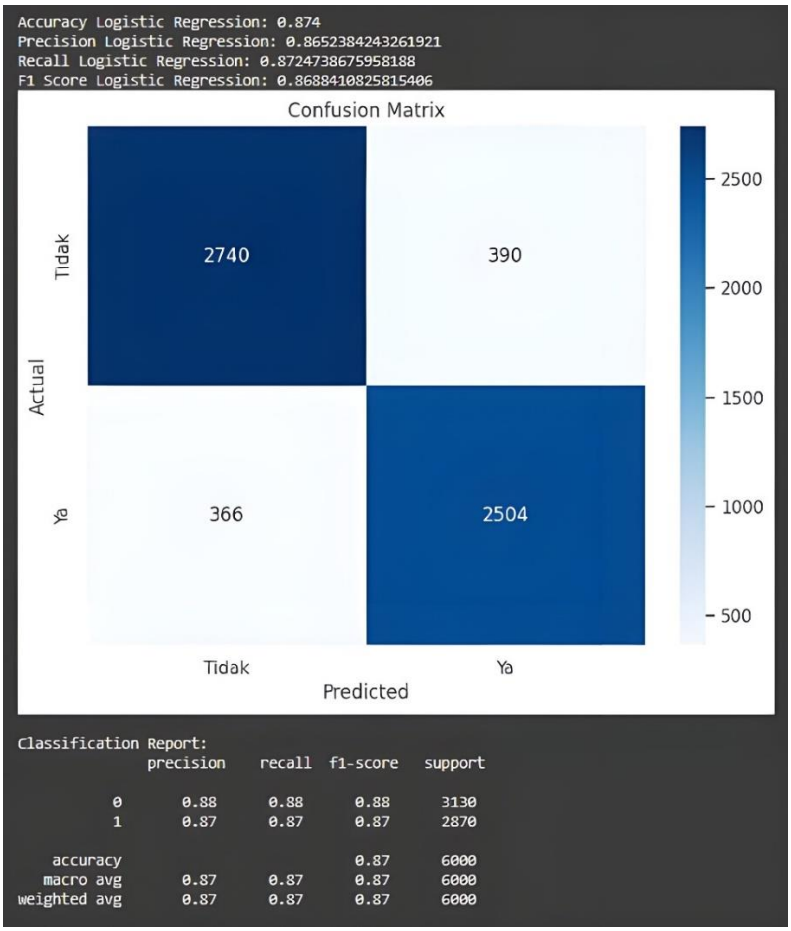
3.4.2 Algoritma XGBoost



Gambar 6 Evaluasi Model XGBoost

Evaluasi model XGBoost pada Gambar 6 menunjukkan akurasi keseluruhan sebesar 94,4%, dengan presisi mencapai 100%, recall 88,38%, dan F1-score 93,83%. Model ini berhasil mengklasifikasikan dengan benar 4.694 sampel pada kelas “Tidak” dan 3.806 sampel pada kelas “Ya”. Tidak ditemukan kesalahan False Positive, tetapi terdapat 500 sampel kelas “Ya” yang salah diklasifikasikan sebagai “Tidak” (False Negative). Precision untuk kelas “Tidak” tercatat di angka 0,90, sementara untuk kelas “Ya” mencapai 1,00, mencerminkan tingkat akurasi tinggi dalam mengenali kelas “Ya”. Recall kelas “Tidak” mencapai 1,00, sedangkan recall untuk kelas “Ya” berada di 0,88, menunjukkan perlunya peningkatan untuk mengurangi jumlah False Negative. F1-score berada pada rentang 0,94 hingga 0,95, menunjukkan keseimbangan yang baik antara precision dan recall. Secara keseluruhan, model ini berkinerja sangat baik, namun masih dapat ditingkatkan dalam hal recall untuk kelas “Ya” guna mengurangi jumlah prediksi False Negative.

3.4.3 Algoritma Logistic Regression



Gambar 7 Evaluasi Model Logistic Regression

Pada evaluasi model Logistic Regression pada Gambar 7, akurasi yang diperoleh sebesar 87,40%, dengan presisi 86,52%, recall 87,24%, dan F1-score 86,88%. Model ini berhasil mengklasifikasikan dengan benar 2.740 sampel untuk kelas "Tidak" dan 2.504 sampel untuk kelas "Ya". Namun, terdapat 390 False Positive dan 366 False Negative. Precision untuk kelas "Tidak" tercatat sebesar 0,88, sedangkan untuk kelas "Ya" adalah 0,87. Recall masing-masing kelas adalah 0,88 dan 0,87, menunjukkan kemampuan deteksi yang cukup baik di kedua kelas. F1-score berkisar antara 0,87 hingga 0,88, yang mencerminkan keseimbangan antara precision dan recall, meskipun ada ruang untuk perbaikan dengan mengurangi kesalahan prediksi.

3.4.4 Hasil Evaluasi Perbandingan Pada Dua Skenario

Hasil evaluasi perbandingan kinerja tiga algoritma dalam dua skenario berbeda, yaitu tanpa Hyperparameter Tuning dan setelah dilakukan Hyperparameter Tuning. Evaluasi ini bertujuan untuk memahami kinerja masing-masing algoritma pada setiap skenario serta melihat sejauh mana pengaruh Hyperparameter Tuning terhadap peningkatan performa algoritma tersebut.

Tabel 1 Perbandingan Evaluasi Model Tanpa Hyperparameter Tuning

Algoritma	Accuracy	Precision	Recall	F1-Score
Random Forest	0.943	1.0	0.8815	0.9370
XGBoost	0.944	1.0	0.8838	0.9383
Logistic Regression	0.874	0.8652	0.8724	0.8688

Tabel 1 membandingkan performa tiga algoritma, Random Forest, XGBoost, dan Logistic Regression berdasarkan akurasi, precision, recall, dan F1-score. Random Forest dan XGBoost menunjukkan akurasi mirip, masing-masing 0,943 dan 0,944, dengan precision sempurna 1,0. XGBoost unggul sedikit dalam recall 0,8838 dibanding Random Forest 0,8815, menunjukkan deteksi positif yang lebih baik. F1-score XGBoost 0,9383 juga sedikit lebih tinggi dibanding Random Forest 0,9370. Logistic Regression memiliki performa lebih rendah pada semua metrik, dengan akurasi 0,874, precision 0.8652, recall 0.8724, dan F1-score 0,868, menunjukkan bahwa ia kurang optimal. Secara keseluruhan, XGBoost menonjol dengan keseimbangan recall dan F1-score yang lebih baik, menjadikannya pilihan lebih andal untuk klasifikasi pada dataset ini.

Tabel 2 Perbandingan Algoritma Setelah Hyperparameter Tuning

Algoritma	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9433	0.9489	0.9433	0.9430
XGBoost	0.9444	0.9498	0.9444	0.9441
Logistic Regression	0.8740	0.8741	0.8740	0.8740

Tabel 2 memperlihatkan kinerja tiga algoritma klasifikasi, XGBoost, Random Forest, dan Logistic Regression, setelah dilakukan hyperparameter tuning, dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score. XGBoost menunjukkan performa tertinggi dengan akurasi 0,9444, precision 0,9498, recall 0,9444, dan F1-score 0,9441, mencerminkan keseimbangan yang optimal dalam mendeteksi kelas positif dan negatif. Random Forest mendekati kinerja XGBoost dengan akurasi 0,9433, precision 0,9489, recall 0,9433, dan F1-score 0,9430, menandakan tuning yang berhasil. Sementara itu, meskipun Logistic Regression telah dituning, performanya tetap lebih rendah dengan semua metrik berada di angka 0,8740. Secara keseluruhan, XGBoost menjadi model terbaik setelah tuning, diikuti oleh Random Forest, sedangkan Logistic Regression kurang optimal untuk dataset ini.

(Revisi) - Perbedaan dari kedua skenario yang ditunjukkan pada Tabel 1 dan Tabel 2 terletak pada peningkatan kinerja algoritma setelah dilakukan tuning hyperparameter. Sebelum Tuning, Random Forest dan XGBoost menunjukkan skor akurasi yang tinggi 0,943 dan 0,944 dengan presisi sempurna, tetapi nilai recall mereka masih kurang. Hal ini mengindikasikan bahwa beberapa sampel positif tidak terdeteksi. Di sisi lain, Logistic Regression memiliki performa yang lebih rendah secara keseluruhan, dengan akurasi 0,874 dan F1-score 0,8688. Setelah tuning, Random Forest dan XGBoost mengalami peningkatan, dengan presisi tetap tinggi, sementara recall membaik, dan menghasilkan F1-score yang lebih tinggi 0,9430 dan 0,9441, yang menunjukkan keseimbangan deteksi yang lebih akurat. Logistic Regression menunjukkan sedikit peningkatan, tetapi performanya tetap tidak sebaik model berbasis pohon keputusan. Temuan ini menyoroti pentingnya tuning hyperparameter dalam meningkatkan akurasi dan mengurangi kesalahan diagnosis, yang sangat penting dalam aplikasi medis. Secara keseluruhan, XGBoost menonjol dengan keseimbangan yang ideal, membuatnya sangat efektif untuk diagnosis dini penyakit paru-paru, sementara Random Forest juga menunjukkan kinerja yang andal, mempertegas potensi besar kedua algoritma ini dalam sistem pendukung keputusan medis.

4 KESIMPULAN DAN SARAN

Berdasarkan evaluasi, XGBoost menonjol sebagai algoritma dengan performa terbaik untuk klasifikasi pada dataset ini, baik sebelum maupun setelah hyperparameter tuning. Sebelum tuning, XGBoost mencatatkan akurasi 94,4%, precision 100%, recall 88,38%, dan F1-score 93,83%. Setelah tuning, performanya meningkat menjadi akurasi 94,44%, precision 94,98%, recall 94,44%, dan F1-score 94,41%, menunjukkan keseimbangan optimal antara precision dan

recall. Random Forest juga menunjukkan hasil mendekati XGBoost dengan akurasi 94,3%, precision 100%, recall 88,15%, dan F1-Score 93,70% sebelum tuning, yang kemudian naik menjadi akurasi 94,33%, precision 94,89%, recall 94,33%, dan F1-Score 94,30% setelah tuning. Sementara itu, Logistic Regression mencatatkan performa lebih rendah, dengan sedikit peningkatan setelah tuning dengan akurasi 87,40%, precision 87,41%, recall 87,40%, dan F1-score 87,40%, menunjukkan keterbatasannya dalam menangani kompleksitas data ini. Hasil ini mengindikasikan bahwa algoritma berbasis pohon keputusan seperti XGBoost dan Random Forest lebih sesuai untuk klasifikasi pada dataset ini karena mampu menjaga akurasi tinggi dan keseimbangan antara precision serta recall.

Penelitian selanjutnya disarankan untuk menguji model ini pada jenis citra medis lain, seperti CT scan atau MRI, untuk mengevaluasi konsistensi performa algoritma pada berbagai tipe data serta memperluas penerapannya dalam diagnosis klinis yang lebih komprehensif.

UCAPAN TERIMA KASIH

Rasa terimakasih kami ucapkan kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Amikom Purwokerto yang mendanai publikasi penelitian ini dan Pusat Studi Sistem Cerdas.

DAFTAR PUSTAKA

[1] L. R. A. Tarigan, "OPTIMALISASI FITUR DENGAN FORWARD SELECTION PADA ESTIMASI TINGKAT PENYAKIT PARU-PARU MENGGUNAKAN ALGORITMA KLASIFIKASI RANDOM FOREST," vol. 8, no. 5, 2024.

[2] R. Harahap, M. Irpan, M. A. Dinata, and L. Efrizoni, "PERBANDINGAN ALGORITMA RANDOM FOREST DAN XGBOOST UNTUK KLASIFIKASI PENYAKIT PARU-PARU BERDASARKAN DATA DEMOGRAFI PASIEN," 2024.

[3] T. K. Abuya, "Lung Cancer Prediction from Elvira Biomedical Dataset Using Ensemble Classifier with Principal Component Analysis," *J. Data Anal. Inf. Process.*, vol. 11, no. 02, pp. 175–199, 2023, doi: 10.4236/jdaip.2023.112010.

[4] U. Chandran, J. Reps, R. Yang, A. Vachani, F. Maldonado, and I. Kalsekar, "Machine Learning and Real-World Data to Predict Lung Cancer Risk in Routine Care," *Cancer Epidemiol. Biomarkers Prev.*, vol. 32, no. 3, pp. 337–343, Mar. 2023, doi: 10.1158/1055-9965.EPI-22-0873.

[5] Dr. M. Kasthuri and M. R. Jency, "Lung Cancer Prediction Using Machine Learning Algorithms on Big Data: Survey," *Int. J. Comput. Sci. Mob. Comput.*, vol. 9, no. 10, pp. 73–77, Oct. 2020, doi: 10.47760/IJCSMC.2020.v09i10.009.

[6] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 1, pp. 114–122, Feb. 2021, doi: 10.29207/resti.v5i1.2880.

[7] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes," *INFORMAL Inform. J.*, vol. 7, no. 3, p. 193, Dec. 2022, doi: 10.19184/isj.v7i3.35441.

[8] Q. Refa Cahyani, M. Januar Finandi, J. Rianti, D. Lestari Arianti, and A. Dwi Pratama Putra, "Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik," *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, doi: 10.55123/jomlai.v1i2.598.

[9] A. D. Azzumzumi, M. Hanafi, and W. M. P. Dhuhita, "Klasifikasi Penyakit Paru-Paru Berdasarkan Peningkatan Kualitas Kontras dan EfficientNet Menggunakan Gambar X-Ray," *Teknika*, vol. 13, no. 2, pp. 293–300, Jul. 2024, doi: 10.34148/teknika.v13i2.881.

[10] A. D. Deva, F. Firdaus, S. Hasyim, B. Yanto, and R. M. Candra, "Klasifikasi Prediksi Penyakit Paru-Paru Normal dengan Pneumonia berdasarkan Citra Image X-ray dengan Optimasi Adam Convolutional Neural Network (CNN)".
[11] S. Goyal and R. Singh, "Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 4, pp. 3239–3259, Apr. 2023, doi: 10.1007/s12652-021-03464-7.
[12] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, p. 652801, Mar. 2021, doi: 10.3389/fenrg.2021.652801.
[13] I. Sen, Md. I. Hossain, Md. F. H. Shakib, Md. A. Imran, and F. Al Faisal, "In Depth Analysis of Lung Disease Prediction Using Machine Learning Algorithms," in *Machine Learning, Image Processing, Network Security and Data Sciences*, vol. 1241, A. Bhattacharjee, S. Kr. Borgohain, B. Soni, G. Verma, and X.-Z. Gao, Eds., in Communications in Computer and Information Science, vol. 1241. , Singapore: Springer Singapore, 2020, pp. 204–213. doi: 10.1007/978-981-15-6318-8_18.
[14] P. Handayani and A. C. Fauzan, "Machine Learning Klasifikasi Status Gizi Balita MenggunakanAlgoritma Random Forest".
[15] F. Diba, "Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, Aug. 2023, doi: 10.33022/ijcs.v12i4.3329.
[16] S. Fan, J. Lin, S. Wu, X. Mu, and J. Guo, "Random forest model can predict the prognosis of hospital-acquired Klebsiella pneumoniae infection as well as traditional logistic regression model," *PLOS ONE*, vol. 17, no. 11, p. e0278123, Nov. 2022, doi: 10.1371/journal.pone.0278123.
[17] Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. RUMPUN Mat. DAN ILMU Pengetah. ALAM*, vol. 2, no. 2, pp. 87–103, Jul. 2023, doi: 10.55606/jurrimipa.v2i2.1470.
[18] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, pp. 21–26, Aug. 2022, doi: 10.31605/jomta.v4i1.1792.
[19] A. F. B. Sajiwo, B. Rahmat, and A. Junaidi, "KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARAN (ISPU) MENGGUNAKAN ALGORITMA XGBOOST DENGAN TEKNIK IMBALANCED DATA (SMOTE)," *J. Inform. Dan Tek. Elektro Terap.*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4699.
[20] M. Syukron, R. Santoso, and T. Widiharih, "PERBANDINGAN METODE SMOTE RANDOM FOREST DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA IMBALANCE CLASS DATA," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, Aug. 2020, doi: 10.14710/j.gauss.v9i3.28915.
[21] A. M. Widodo, Y. S. Anggraeni, N. Anwar, A. Ichwani, and B. Anggara, "Performansi K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes," 2021.
[22] Y. N. Khoiril Umat, D. Rusyda Nafsyi, D. Kusumaningsih, and L. Hakim, "ANALISIS FAKTOR YANG MEMPENGARUHI PEMILIHAN GUBERNUR DAERAH KHUSUS JAKARTA MENGGUNAKAN ALGORITMA NAIVE BAYES DAN REGRESI LOGISTIK," *Rabit J. Teknol. Dan Sist. Inf. Univrab*, vol. 9, no. 2, pp. 211–224, Jul. 2024, doi: 10.36341/rabit.v9i2.4778.
[23] Dr. M. Kasthuri and M. R. Jency, "Improving the Performance of Lung Cancer Prediction Using Machine Learning Techniques on Big Data," *Int. J. Comput. Sci. Mob. Comput.*, vol. 9, no. 10, pp. 64–72, Oct. 2020, doi: 10.47760/IJCSMC.2020.v09i10.008.
[24] S. Sujana, A. R. Juwita, and S. Faisal, "Penerapan Metode Regresi Logistik Untuk Memprediksi Peristiwa Biner Pasien Pasca Operasi Kanker Payudara," vol. 5, no. 4, 2024.

[25] K. Shankar, Y. Zhang, Y. Liu, L. Wu, and C.-H. Chen, "Hyperparameter Tuning Deep Learning for Diabetic Retinopathy Fundus Image Classification," *IEEE Access*, vol. 8, pp. 118164–118173, 2020, doi: 10.1109/ACCESS.2020.3005152.
[26] "Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 6, pp. 198–207, Dec. 2021, doi: 10.22266/ijies2021.1231.19.
[27] T. Huang, D. Le, L. Yuan, S. Xu, and X. Peng, "Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit," *PLOS ONE*, vol. 18, no. 1, p. e0280606, Jan. 2023, doi: 10.1371/journal.pone.0280606.
[28] J. Anggraini and D. Alita, "Implementasi Metode SVM Pada Sentimen Analisis Terhadap Pemilihan Presiden (Pilpres) 2024 Di Twitter," *J. Inform. J. Pengemb. IT*, vol. 9, no. 2, pp. 102–111, Aug. 2024, doi: 10.30591/jpit.v9i2.6560.