Optimasi dan Evaluasi Komparatif Algoritma Klasifikasi untuk Prediksi Risiko Gangguan Tidur Berbasis Data Gaya Hidup

Septi Isdayanna[1], Asmunin[2]

Program Studi Manajemen Informatika Universitas Negeri Surabaya
Jl. Ketintang, Kec. Gayungan, Kota Surabaya, Universitas Negeri Surabaya
[1]septi.22045@mhs.unesa.ac.id
[2]asmunin@unesa.ac.id

Abstrak *- Evaluasi komparatif algoritma machine learning pada data kesehatan tabular sering berfokus pada satu metrik performa tanpa prosedur yang konsisten, sehingga hasil perbandingan berpotensi kurang valid. Penelitian ini menyusun dan menerapkan kerangka evaluasi komparatif untuk menguji performa tiga algoritma klasifikasi tunggal, yaitu Logistic Regression, Random Forest, dan Gradient Boosting, dalam memprediksi risiko gangguan tidur (Healthy, Insomnia, Sleep Apnea) menggunakan Sleep Health and Lifestyle Dataset. Kerangka evaluasi mencakup audit data yang mengeliminasi 67,65% observasi duplikat, pipeline pra-pemrosesan yang identik, serta optimasi hyperparameter melalui Grid Search dengan Stratified 5-Fold Cross-Validation. Random Forest (Tuned) memberikan performa keseluruhan terbaik (Accuracy 91,56%; Macro F1-Score 0,9100), sedangkan Gradient Boosting memperoleh Macro ROC-AUC tertinggi (0,9815); Logistic Regression tidak memenuhi ambang akurasi minimum yang ditetapkan (84,14%). Temuan ini menunjukkan bahwa kualitas evaluasi komparatif pada data kesehatan tabular tidak hanya ditentukan oleh pemilihan algoritma, tetapi juga oleh konsistensi prosedur evaluasi yang digunakan.*

Kata kunci: *Evaluasi Komparatif, Klasifikasi Gangguan Tidur, Optimasi Hyperparameter, Data Kesehatan Tabular, Machine Learning*

Abstract *- Comparative evaluation of machine learning algorithms on tabular health data often relies on a single performance metric without a consistent procedure, potentially compromising the validity of the comparison. This study designs and applies a comparative evaluation framework to examine the performance of three single classification algorithms, specifically Logistic Regression, Random Forest, and Gradient Boosting, in predicting sleep disorder risk (Healthy, Insomnia, Sleep Apnea) using the Sleep Health and Lifestyle Dataset. The framework comprises a data audit that eliminated 67.65% duplicate observations, an identical preprocessing pipeline , and hyperparameter optimization via Grid Search with Stratified 5-Fold Cross-Validation. Random Forest (Tuned) achieved the best overall performance (91.56% accuracy; 0.9100 Macro F1-Score), while Gradient Boosting obtained the highest Macro ROC-AUC (0.9815); Logistic Regression failed to meet the established minimum accuracy threshold (84.14%). These findings indicate that the quality of comparative evaluation on tabular health data is determined not only by algorithm selection, but also by the consistency of the evaluation procedure applied.*

Keywords: *Comparative Evaluation, Sleep Disorder Classification, Hyperparameter Optimization, Tabular Health Data, Machine Learning*

I. PENDAHULUAN

Berdasarkan indikator *Life's Essential 8* dari *American Heart Association* [1], tidur yang restoratif diakui sebagai salah satu pilar utama kesehatan kardiovaskular, pemulihan kognitif, dan kestabilan emosional [2]. Data nasional turut mengonfirmasi urgensi ini: prevalensi *Insomnia* pada populasi dewasa Indonesia dilaporkan mencapai 43,7% [3], sebuah angka yang menegaskan bahwa persoalan ini bukan sekadar isu gaya hidup individual, melainkan tantangan kesehatan masyarakat berskala nasional yang menuntut instrumen penapisan dini yang presisi, efisien, dan berskala massal.

Perkembangan *machine learning* mendorong semakin luasnya pemanfaatan data kesehatan tabular sebagai dasar pengembangan model prediksi risiko gangguan tidur. Sebagai instrumen penapisan skala luas, pendekatan berbasis rekam data gaya hidup tabular, yang mencakup durasi tidur, tingkat stres, metrik aktivitas, dan fisiologis dasar, menawarkan alternatif yang lebih terjangkau, karena dapat dikumpulkan secara mandiri oleh pengguna tanpa memerlukan sensor biologis khusus yang relatif mahal dan sulit diakses secara luas [4]. Kendati demikian, pemodelan berbasis rekam data tabular menghadapi tantangan struktural tersendiri. Karakteristik data ini secara inheren memiliki relasi non-linear yang kompleks serta heterogenitas subjek yang tinggi [5]. Di samping itu, distribusi data medis sering kali berada dalam kondisi tidak seimbang antarkelas (*imbalanced data*), yang menuntut penanganan algoritmik secara khusus agar model tidak bias terhadap kelas mayoritas [6].

Meskipun evaluasi algoritma *machine learning* pada *dataset* gangguan tidur telah banyak dilakukan, literatur terdahulu kerap menunjukkan variasi performa yang signifikan. Sebagai ilustrasi, salah satu penelitian yang membandingkan beberapa algoritma klasifikasi pada *Sleep Health and Lifestyle Dataset* melaporkan rentang akurasi sebesar 87,50–94,44% [7]. Perbedaan performa yang dilaporkan tidak hanya dipengaruhi oleh karakteristik algoritma itu sendiri, melainkan juga oleh perbedaan protokol evaluasi yang digunakan, di mana penelitian tersebut cenderung mengekspos akurasi tertinggi tanpa merinci proses audit kualitas data, pembersihan anomali, maupun skema optimalisasi parameter (*hyperparameter tuning*) yang diterapkan secara seragam pada seluruh model yang dibandingkan. Berbeda dengan penelitian terdahulu yang

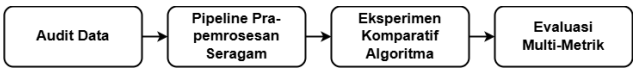
berfokus pada pelaporan performa algoritma, penelitian ini menitikberatkan pada validitas proses evaluasi melalui penerapan protokol eksperimen yang konsisten, sehingga perbandingan performa antaralgoritma dapat dilakukan secara lebih objektif.

Ketiadaan standardisasi evaluasi ini menyebabkan munculnya dua celah metodologis yang berisiko. Pertama, keberadaan observasi duplikat pada *dataset* publik dapat memicu kebocoran data (*data leakage*). Kondisi ini menghasilkan estimasi performa yang terlalu optimistis dan tidak mencerminkan kemampuan generalisasi model pada data baru, yang mana isu ini diakui sebagai penyumbang utama krisis reproduktibilitas (*reproducibility crisis*) pada sains kecerdasan buatan [8]. Kedua, ketergantungan absolut pada metrik Akurasi dapat menyembunyikan paradoks evaluasi. Pada data yang tidak seimbang, akurasi tinggi kerap menutupi tingginya angka *False Negative* pada kelas minoritas, padahal dalam konteks klinis, kegagalan mendeteksi kelainan (seperti *Sleep Apnea*) jauh lebih fatal dibandingkan alarm palsu [6]. Oleh karena itu, evaluasi komprehensif menggunakan metrik pendukung seperti *Macro F1-Score*, *Recall*, dan *ROC-AUC* perlu dilakukan.

Menjawab celah metodologis tersebut, penelitian ini bertujuan menyusun dan menerapkan kerangka evaluasi komparatif yang sistematis untuk menguji tiga algoritma klasifikasi tunggal, yaitu *Logistic Regression* [9], *Random Forest* [10], dan *Gradient Boosting* [11], melalui audit deduplikasi data secara eksplisit, penerapan *pipeline* pra-pemrosesan yang identik, serta optimasi *hyperparameter* berbasis *Grid Search* dengan *Stratified 5-Fold Cross-Validation*, disertai evaluasi performa menggunakan metrik yang komprehensif. Kontribusi utama penelitian ini bukan hanya membandingkan performa ketiga algoritma tersebut, melainkan menunjukkan bahwa validitas evaluasi pada data kesehatan tabular sangat dipengaruhi oleh kualitas data yang telah diaudit, konsistensi *pipeline* eksperimen, keseragaman proses optimasi, dan kelengkapan metrik evaluasi yang digunakan, sehingga kesimpulan performa yang dihasilkan lebih objektif dan dapat direproduksi. Ruang lingkup penelitian ini difokuskan pada tahap pengembangan dan evaluasi model prediktif berbasis algoritma klasifikasi tunggal; pengembangan komponen lanjutan di luar proses prediksi tidak menjadi bagian dari cakupan penelitian ini.

II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif melalui eksperimen komputasional berbasis *supervised learning* pada *Sleep Health and Lifestyle Dataset*. Untuk menjamin bahwa perbandingan performa antaralgoritma murni mencerminkan kapasitas intrinsik masing-masing model dan bukan artefak dari perbedaan kondisi eksperimen, penelitian ini disusun mengikuti kerangka evaluasi komparatif, sebagaimana ditunjukkan pada Gambar 1: (A) audit data untuk memastikan validitas *dataset* sebelum pemodelan, (B) penerapan *pipeline* pra-pemrosesan yang identik pada seluruh algoritma, (C) eksperimen komparatif algoritma melalui skenario baseline dan optimasi hyperparameter yang setara, dan (D) evaluasi performa menggunakan metrik yang komprehensif. Keempat tahap tersebut dirancang sebagai satu kesatuan prosedur evaluasi yang saling bergantung, sehingga perubahan pada salah satu tahap berpotensi memengaruhi validitas perbandingan performa antaralgoritma; oleh karena itu, setiap tahap diterapkan secara konsisten pada seluruh algoritma yang dibandingkan.



Gambar. 1 Kerangka Evaluasi Komparatif Algoritma Klasifikasi

A. Dataset dan Audit Data

Penelitian ini menggunakan *Sleep Health and Lifestyle Dataset* yang tersedia secara publik di *platform* Kaggle, terdiri atas 15.000 observasi dengan 13 kolom, mencakup satu variabel identitas, sebelas atribut gaya hidup dan fisiologis (di antaranya durasi tidur, tingkat stres, aktivitas fisik, indeks massa tubuh, tekanan darah, dan detak jantung), serta satu variabel target klasifikasi tiga kelas (*Healthy*, *Insomnia*, *Sleep Apnea*). *Dataset* ini bersifat sintetis, sehingga sesuai digunakan untuk validasi kerangka evaluasi tanpa terkendala isu privasi data medis riil. Penggunaan dataset sintetis memungkinkan penelitian difokuskan pada validitas prosedur evaluasi tanpa dipengaruhi keterbatasan akses terhadap data medis riil maupun isu etika penggunaan data pasien.

Audit data diawali dengan pemeriksaan struktur, distribusi kelas, dan potensi anomali. Audit dilakukan melalui pemeriksaan nilai identik pada seluruh atribut prediktor dan label setelah variabel identitas dihapus, yang mengungkap 10.148 baris (67,65%) identik pada seluruh atribut. Observasi duplikat tersebut dieliminasi secara eksplisit sebagai bagian dari audit untuk mencegah kebocoran data pada tahap evaluasi, menyisakan 4.852 observasi unik dengan distribusi kelas yang berubah menjadi tidak seimbang (*Healthy* 51,13%; *Insomnia* 24,67%; *Sleep Apnea* 24,20%). Fitur komposit yang memuat dua nilai sekaligus, seperti tekanan darah, diekstraksi menjadi variabel numerik terpisah (sistolik dan diastolik), sementara kategori dengan redundansi semantik dikonsolidasikan.

B. Pipeline Pra-pemrosesan Seragam

Dataset hasil audit dipartisi menjadi data latih dan data uji dengan rasio 80:20 menggunakan *stratified sampling* (3.881 data latih, 971 data uji), dilakukan sebelum tahap transformasi fitur untuk mencegah kebocoran informasi statistik dari data uji ke data latih. Seluruh fitur numerik distandardisasi, fitur kategorikal nominal di-*encode* melalui one-hot encoding, dan fitur berjenjang menggunakan *ordinal encoding*; seluruh transformasi tersebut disatukan dalam satu objek *pipeline* pra-pemrosesan tunggal yang diterapkan secara identik pada ketiga algoritma yang dibandingkan. Keseragaman *pipeline* ini merupakan komponen penting dalam kerangka evaluasi yang diusulkan, untuk memastikan bahwa perbedaan performa yang teramati tidak dipengaruhi oleh perbedaan perlakuan pra-pemrosesan antaralgoritma.

C. Eksperimen Komparatif Algoritma

Tiga algoritma klasifikasi tunggal dibandingkan dalam penelitian ini, yaitu *Logistic Regression* [9], *Random Forest* [10], dan *Gradient Boosting* [11]. Ketiganya dipilih untuk

merepresentasikan pendekatan klasifikasi yang berbeda: *Logistic Regression* sebagai model linear yang memberikan baseline dengan varians rendah; *Random Forest* yang menggabungkan prediksi banyak pohon keputusan yang dilatih pada subset data acak untuk mereduksi varians; dan *Gradient Boosting* yang membangun rangkaian pohon keputusan secara bertahap, dengan setiap pohon baru mengoreksi kesalahan pohon sebelumnya, untuk mereduksi bias.

Untuk memastikan bahwa perbandingan performa antaralgoritma dilakukan pada kondisi yang setara dan tidak dipengaruhi oleh perbedaan tingkat optimasi, setiap algoritma dievaluasi melalui dua skenario: skenario baseline menggunakan parameter bawaan (*default*), dan skenario optimasi menggunakan *Grid Search* dengan target metrik *Macro F1-Score*. Pemilihan *Macro F1-Score* sebagai target optimasi didasarkan pada karakteristik distribusi kelas yang tidak seimbang. Optimasi ini divalidasi melalui *Stratified 5-Fold Cross-Validation* [12] guna menjamin konsistensi proporsi kelas target pada setiap lipatan data latih. Kedua skenario tersebut dibandingkan untuk mengevaluasi pengaruh penerapan optimasi yang setara terhadap validitas hasil komparasi antaralgoritma, sehingga interpretasi performa tidak dipengaruhi oleh perbedaan tingkat optimasi model.

Penanganan ketidakseimbangan kelas dilakukan melalui pembobotan kelas otomatis (class_weight='balanced') pada *Logistic Regression* dan *Random Forest*. Berbeda dengan *Logistic Regression* dan *Random Forest* yang menggunakan pembobotan kelas otomatis, implementasi *Gradient Boosting* pada penelitian ini tidak menerapkan mekanisme pembobotan kelas sehingga model dilatih menggunakan distribusi data hasil audit sebagaimana adanya. Konfigurasi hyperparameter optimal yang dihasilkan adalah *Logistic Regression* (C=1, penalty=L1, solver=saga), *Random Forest* (n_estimators=200, max_depth=None, min_samples_split=10), dan *Gradient Boosting* (learning_rate=0,2, n_estimators=200, max_depth=3, subsample=1,0).

D. Evaluasi Multi-Metrik

Performa klasifikasi dievaluasi menggunakan lima metrik, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*, dengan penekanan pada *macro-averaged F1-score* untuk menjamin keadilan evaluasi pada kelas minoritas berisiko klinis (*Insomnia* dan *Sleep Apnea*). Stabilitas generalisasi masing-masing model diverifikasi dengan membandingkan skor validasi silang (*CV F1-Macro*) terhadap skor pada data uji independen, guna mengonfirmasi ketiadaan indikasi *overfitting*. Analisis *confusion matrix* turut dilakukan untuk mengidentifikasi pola kesalahan klasifikasi, khususnya kegagalan mendeteksi kondisi *Sleep Apnea* sebagai *false negative*. Kriteria keberhasilan ditetapkan sejak tahap perancangan, mencakup ambang minimum akurasi ≥85%, *Macro F1-Score* dan *Macro Recall* masing-masing ≥0,800, serta *Macro ROC-AUC* ≥0,850. Ambang tersebut ditetapkan sebagai target evaluasi penelitian berdasarkan karakteristik permasalahan klasifikasi multikelas pada domain kesehatan, dan tidak dimaksudkan sebagai standar klinis universal.

III. HASIL PENELITIAN DAN PEMBAHASAN

Bagian ini menyajikan hasil evaluasi komparatif tiga algoritma klasifikasi tunggal, yaitu *Logistic Regression*, *Random Forest*, dan *Gradient Boosting*, yang dilatih dan diuji mengikuti kerangka evaluasi yang diuraikan pada Bagian II. Evaluasi dilakukan pada 971 data uji hasil partisi stratifikasi 80:20, dengan distribusi kelas aktual yang identik untuk seluruh model (*Healthy* 496, *Insomnia* 240, *Sleep Apnea* 235), sehingga performa antaralgoritma dapat dibandingkan secara langsung tanpa bias akibat perbedaan komposisi data.

A. Hasil Evaluasi Komparatif Model

Optimasi *hyperparameter* melalui *Grid Search* dengan *Stratified 5-Fold Cross-Validation* terbukti meningkatkan performa ketiga model dibandingkan konfigurasi *baseline* (Tabel I), dengan peningkatan *Macro F1-Score* paling signifikan pada *Random Forest* (dari 0,8774 menjadi 0,9100) dan *Gradient Boosting* (dari 0,8788 menjadi 0,9033), sementara *Logistic Regression* menunjukkan peningkatan marjinal (dari 0,8302 menjadi 0,8316). Peningkatan ini menunjukkan bahwa penerapan prosedur optimasi yang identik pada seluruh algoritma berhasil mengurangi pengaruh perbedaan konfigurasi awal, sehingga perbandingan performa pada Tabel I dan Tabel II dilakukan pada kondisi evaluasi yang setara, bukan dipengaruhi oleh disparitas upaya *tuning* antaralgoritma.

TABEL I
PERBANDINGAN PERFORMA LINTAS KONFIGURASI MODEL

Model	Konfigurasi	*Test Acc.*	*Test F1-Macro*
Logistic Regression	*Tuned*	0,8414	0,8316
	Baseline	0,8404	0,8302
Random Forest	*Tuned*	0,9156	0,9100
	Baseline	0,8857	0,8774
Gradient Boosting	*Tuned*	0,9094	0,9033
	Baseline	0,8877	0,8788

Stabilitas generalisasi ketiga model turut diverifikasi melalui perbandingan skor validasi silang (*CV F1-Macro*) terhadap skor pengujian aktual, dengan kesenjangan yang tercatat positif dan sangat kecil pada seluruh model (*Logistic Regression* +0,0032; *Random Forest* +0,0169; *Gradient Boosting* −0,0004), mengindikasikan bahwa proses *tuning* tidak menyebabkan *overfitting* terhadap data latih.

Terhadap kriteria keberhasilan yang telah ditetapkan, yaitu akurasi ≥85%, Macro F1-Score dan Macro Recall ≥0,800, serta Macro ROC-AUC ≥0,850, Random Forest dan Gradient Boosting berhasil memenuhi seluruh ambang batas yang ditentukan (Tabel II). Sebaliknya, Logistic Regression hanya memenuhi tiga dari empat kriteria karena nilai akurasinya sebesar 0,8414 masih sedikit berada di bawah target minimum 85%. Temuan ini menunjukkan bahwa kedua algoritma berbasis pohon lebih efektif dalam menghasilkan performa klasifikasi yang konsisten pada dataset gangguan tidur yang digunakan.

TABEL II
RINGKASAN PERFORMA MODEL TERBAIK (*TUNED*)

Model	Accuracy	Macro F1-Score	Macro Recall	Macro ROC-AUC
Logistic Regression	0,8414	0,8316	0,8438	0,9496
Random Forest	0,9156	0,9100	0,9230	0,9767
Gradient Boosting	0,9094	0,9033	0,9124	0,9815

Tabel II turut menunjukkan bahwa tidak terdapat satu algoritma yang secara absolut mendominasi seluruh metrik evaluasi: *Random Forest* unggul pada *Accuracy*, *Macro F1-Score*, dan *Macro Recall*, sedangkan *Gradient Boosting* justru mencatatkan *Macro ROC-AUC* tertinggi (0,9815 berbanding 0,9767). Perbedaan ini menunjukkan bahwa interpretasi performa sangat dipengaruhi oleh metrik evaluasi yang dipilih, sehingga penggunaan satu metrik saja, termasuk *Accuracy* yang paling umum digunakan, berpotensi menghasilkan kesimpulan yang tidak lengkap. Oleh karena itu, pemilihan algoritma pada konteks aplikatif perlu mempertimbangkan tujuan evaluasi dan karakteristik metrik yang relevan, bukan hanya satu indikator performa tunggal.

B. *Analisis Performa Per Kelas*

TABEL III
CLASSIFICATION REPORT KOMPARATIF (*TUNED*)

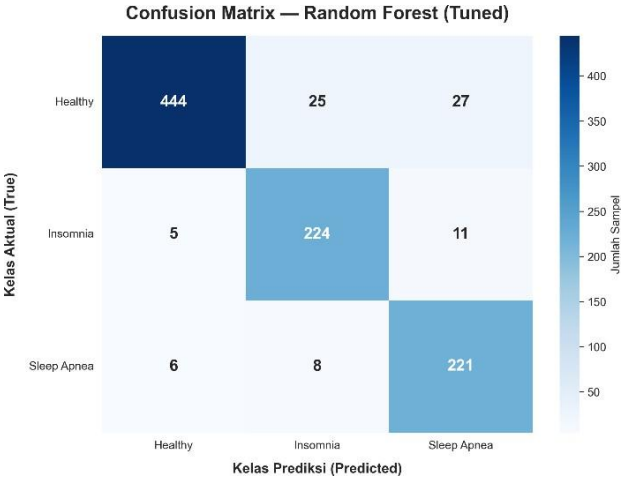
Model	Kelas	Precision	Recall	F1-Score	Support
Logistic Regression	Healthy	0,9180	0,8347	0,8743	496
	Insomnia	0,7621	0,8542	0,8055	240
	Sleep Apnea	0,7888	0,8426	0,8148	235
	Macro Avg	0,8230	0,8438	0,8316	971
Random Forest	Healthy	0,9758	0,8952	0,9338	496
	Insomnia	0,8716	0,9333	0,9014	240
	Sleep Apnea	0,8533	0,9404	0,8947	235
	Macro Avg	0,9002	0,9230	0,9100	971
Gradient Boosting	Healthy	0,9572	0,9012	0,9283	496
	Insomnia	0,8645	0,9042	0,8839	240
	Sleep Apnea	0,8656	0,9319	0,8975	235
	Macro Avg	0,8958	0,9124	0,9033	971

Rincian performa per kelas menunjukkan pola yang konsisten pada ketiga model: *precision* tertinggi selalu diperoleh pada kelas *Healthy*, sementara *recall*-nya justru yang terendah di antara ketiga kelas. Pola ini mengindikasikan bahwa seluruh model relatif konservatif sebelum melabeli seseorang sebagai sehat, namun sensitif dalam mendeteksi kelas gangguan tidur, sebuah karakteristik yang sesuai dengan kebutuhan konteks penapisan dini.

Namun demikian, magnitudo sensitivitas ini bervariasi cukup besar antaralgoritma. Pada kelas *Sleep Apnea*, *recall Random Forest* (0,9404) dan *Gradient Boosting* (0,9319) jauh melampaui *Logistic Regression* (0,8426) dengan selisih hingga sembilan poin persentase, yang berarti *Logistic Regression* melewatkan hampir 16% kasus *Sleep Apnea* aktual (37 dari 235 sampel), dibandingkan sekitar 6% pada kedua model berbasis pohon. Kesenjangan ini tidak sepenuhnya tercermin pada selisih *Accuracy* antarmodel yang tampak lebih moderat (Tabel I). Temuan ini menegaskan alasan pemilihan *Macro F1-Score* sebagai metrik utama pengoptimalan pada tahap Metode: karena setiap kelas diberi bobot yang sama tanpa memperhitungkan proporsi sampel, *Macro F1-Score* mampu menangkap kelemahan model pada kelas minoritas seperti yang teramati pada *Logistic Regression*, sekalipun *Accuracy* keseluruhannya masih tampak kompetitif (84,14%).

C. *Analisis Kesalahan Klasifikasi Random Forest (Tuned)*

Sebagai model dengan performa keseluruhan terbaik pada penelitian ini, analisis kesalahan klasifikasi lebih lanjut difokuskan pada *Random Forest* (*Tuned*). Gambar 2 menunjukkan confusion matrix pada 971 data uji.



Gambar. 2 Confusion Matrix *Random Forest* (*Tuned*)

Dari 971 sampel uji, model gagal mendeteksi 16 dari 240 kasus *Insomnia* aktual dan 14 dari 235 kasus *Sleep Apnea* aktual sebagai *false negative*. Kesalahan yang muncul juga menunjukkan pola tumpang tindih antara kedua gangguan tidur itu sendiri, dengan 11 kasus *Insomnia* diprediksi sebagai *Sleep Apnea* dan 8 kasus sebaliknya. Pola ini konsisten dengan adanya kemiripan karakteristik antarkedua kelas pada *dataset*, sehingga pemisahan keduanya menjadi lebih menantang bagi model klasifikasi manapun.

Perbandingan pola kesalahan yang sama pada ketiga model turut memperkuat interpretasi ini: total kesalahan silang antara *Insomnia* dan *Sleep Apnea* tercatat 19 kasus pada *Random Forest*, 19 kasus pada *Gradient Boosting*, namun meningkat hampir dua kali lipat menjadi 35 kasus pada *Logistic Regression*. Pola ini mengindikasikan bahwa batas keputusan non-linear yang ditangkap oleh model berbasis pohon lebih mampu memisahkan kedua kelas dengan karakteristik tumpang tindih tersebut dibandingkan batas keputusan linear, konsisten dengan kemampuan model berbasis pohon dalam merepresentasikan hubungan non-linear pada data tabular.

D. Pembahasan

Secara keseluruhan, temuan ini menjawab langsung celah metodologis yang diidentifikasi pada Bagian I. Audit deduplikasi eksplisit terhadap 67,65% observasi duplikat mencegah estimasi performa yang optimistis akibat kebocoran data, sementara *pipeline* pra-pemrosesan dan skema optimasi yang identik memastikan bahwa selisih performa yang teramati mencerminkan kapasitas intrinsik algoritma, bukan artefak perlakuan eksperimen. Sebagai perbandingan, akurasi *Random Forest tuned* pada penelitian ini (91,56%) berada dalam rentang yang dilaporkan studi sejenis sebelumnya (87,50-94,44%) [7], namun diperoleh melalui prosedur yang lebih transparan dan dapat direproduksi.

Perbedaan besaran peningkatan performa akibat tuning turut mengonfirmasi karakteristik masing-masing paradigma algoritma. *Random Forest* dan *Gradient Boosting*, sebagai model berbasis pohon, memiliki ruang hyperparameter yang lebih memengaruhi kompleksitas model (seperti jumlah estimator dan kedalaman pohon), sehingga penyesuaian parameter tersebut mampu menangkap relasi non-linear pada data secara lebih optimal dibandingkan konfigurasi default. Sebaliknya, peningkatan marjinal pada *Logistic Regression* konsisten dengan sifat dasarnya sebagai model linear: penyesuaian kekuatan regularisasi dan jenis *solver* dapat mengoptimalkan batas keputusan linear yang ada, tetapi tidak mengubah kapasitas model untuk mempelajari relasi non-linear pada data. Menariknya, capaian *recall Gradient Boosting* pada kedua kelas minoritas (*Insomnia* 0,9042; *Sleep Apnea* 0,9319) sebanding dengan *Random Forest* meski tanpa mekanisme pembobotan kelas eksplisit, mengindikasikan bahwa sifat sekuensial model dalam mengoreksi kesalahan secara umum turut memberikan sensitivitas yang memadai terhadap kelas minoritas pada kasus ini.

Keunggulan *Gradient Boosting* pada *Macro ROC-AUC* meski tertinggal tipis pada metrik lain kemungkinan berkaitan dengan mekanisme pembelajarannya yang bersifat sekuensial, di mana setiap pohon baru secara eksplisit mengoreksi residual kesalahan pohon sebelumnya, berpotensi menghasilkan estimasi probabilitas yang lebih terurut (*rank-consistent*) di seluruh rentang ambang keputusan dibandingkan agregasi rata-rata pada *Random Forest*. Namun demikian, penelitian ini tidak melakukan pengujian kalibrasi probabilitas secara langsung (seperti *reliability* diagram atau *Brier score*), sehingga interpretasi ini bersifat plausibel berdasarkan karakteristik algoritmik dan belum dapat dipastikan sebagai penyebab tunggal tanpa pengujian lebih lanjut.

Evaluasi multi-metrik yang diterapkan membuktikan tesis utama penelitian ini: pemilihan "algoritma terbaik" tidak dapat disandarkan pada satu metrik tunggal. Pola-pola ini, yaitu gap *recall* pada kelas minoritas, perbedaan peringkat antarmetrik, serta sensitivitas berbeda terhadap tuning, tidak akan sepenuhnya terungkap apabila evaluasi hanya bersandar pada *Accuracy* sebagaimana banyak dilakukan pada penelitian terdahulu.

Temuan-temuan tersebut secara kolektif menunjukkan bahwa validitas evaluasi komparatif pada data kesehatan tabular tidak semata ditentukan oleh algoritma mana yang dipilih, melainkan oleh konsistensi kerangka evaluasi yang menaunginya, yang mencakup audit kualitas data sebelum pemodelan, keseragaman *pipeline* pra-pemrosesan, kesetaraan proses optimasi, hingga kelengkapan metrik yang digunakan untuk menilai hasil. Keempat komponen ini saling bergantung: audit data mencegah estimasi performa yang optimistis akibat kebocoran, *pipeline* dan optimasi yang seragam memastikan selisih performa mencerminkan kapasitas algoritma dan bukan artefak eksperimen, sementara evaluasi multi-metrik mengungkap nuansa performa per kelas dan per metrik yang tersembunyi di balik *Accuracy* semata. Dengan demikian, kesimpulan performa yang dihasilkan pada Bagian III.A hingga III.C bukan sekadar produk dari tiga algoritma yang diuji, melainkan produk dari kerangka evaluasi yang diterapkan secara konsisten terhadap ketiganya.

Meskipun demikian, sejumlah keterbatasan perlu dicermati. Pertama, seluruh evaluasi dilakukan pada dataset sintetis, sehingga generalisasi temuan terhadap data klinis riil masih memerlukan validasi lebih lanjut. Kedua, penelitian ini murni berfokus pada perbandingan performa prediktif; kerangka evaluasi yang diusulkan belum diuji pada dataset kesehatan tabular lain di luar konteks gangguan tidur, sehingga generalisasi kerangka ini ke domain lain masih bersifat hipotesis yang perlu diuji pada penelitian selanjutnya.

IV. KESIMPULAN

Penelitian ini berhasil menyusun dan menerapkan kerangka evaluasi komparatif untuk menguji tiga algoritma klasifikasi tunggal, yaitu *Logistic Regression*, *Random Forest*, dan *Gradient Boosting*, dalam memprediksi risiko gangguan tidur berbasis data gaya hidup tabular, melalui audit data, *pipeline* pra-pemrosesan yang identik, dan optimasi *hyperparameter* yang setara pada seluruh algoritma yang dibandingkan.

Hasil evaluasi menunjukkan *Random Forest* (*Tuned*) sebagai model dengan performa keseluruhan terbaik (*Accuracy* 91,56%, *Macro F1-Score* 0,9100, *Macro Recall* 0,9230), diikuti *Gradient Boosting* yang justru mencatatkan *Macro ROC-AUC* tertinggi (0,9815), sementara *Logistic Regression* tidak memenuhi ambang akurasi minimum yang ditetapkan (84,14% dari target $\geqslant$85%) dan menunjukkan kelemahan pada *recall* kelas minoritas berisiko klinis.

Kontribusi utama penelitian ini bukan terletak pada identifikasi algoritma dengan performa tertinggi semata, melainkan pada bukti empiris bahwa validitas evaluasi komparatif pada data kesehatan tabular dipengaruhi oleh konsistensi kerangka evaluasi yang menaunginya. Temuan bahwa tidak terdapat satu algoritma yang secara absolut mendominasi seluruh metrik, di mana *Random Forest* unggul pada *Accuracy* dan *Macro F1-Score*, namun *Gradient Boosting* unggul pada *Macro ROC-AUC*, menegaskan bahwa evaluasi berbasis metrik tunggal berisiko menghasilkan kesimpulan yang tidak lengkap. Secara keseluruhan, hasil ini menunjukkan bahwa audit data, *pipeline* pra-pemrosesan yang seragam, optimasi yang setara, dan evaluasi multi-metrik merupakan komponen yang saling melengkapi dalam menghasilkan evaluasi komparatif yang lebih valid dibandingkan apabila salah satu tahapan tersebut diabaikan.

Penelitian ini menunjukkan bahwa kerangka evaluasi yang diterapkan mampu menghasilkan proses evaluasi komparatif

yang konsisten pada kasus klasifikasi risiko gangguan tidur berbasis data kesehatan tabular, dan berpotensi dijadikan acuan metodologis bagi penelitian serupa pada domain kesehatan tabular lainnya. Dengan demikian, penelitian ini menunjukkan bahwa kualitas evaluasi komparatif pada *machine learning* tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh konsistensi prosedur evaluasi yang mendasarinya.

Penelitian ini memiliki keterbatasan yang membuka ruang pengembangan lanjutan, terutama pada validasi menggunakan data klinis riil dan pengujian kerangka evaluasi pada domain kesehatan tabular lain di luar konteks gangguan tidur.

REFERENSI

[1] D. M. Lloyd-Jones *et al.*, "Life's Essential 8: Updating and Enhancing the American Heart Association's Construct of Cardiovascular Health: A Presidential Advisory from the American Heart Association," Aug. 02, 2022, *Lippincott Williams and Wilkins*. doi: 10.1161/CIR.0000000000001078.

[2] A. Hyndych, R. El-Abassi, and E. C. Mader, "The Role of Sleep and the Effects of Sleep Loss on Cognitive, Affective, and Behavioral Processes," *Cureus*, May 2025, doi: 10.7759/cureus.84232.

[3] H. Edison and O. Nainggolan, "Hubungan Insomnia dengan Hipertensi," *Buletin Penelitian Sistem Kesehatan*, vol. 24, no. 1, pp. 46–56, Feb. 2021, doi: 10.22435/hsr.v24i1.3579.

[4] A. Zinzuwadia and J. P. Singh, "Wearable devices—addressing bias and inequity," Dec. 01, 2022, *Elsevier Ltd*. doi: 10.1016/S2589-7500(22)00194-7.

[5] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is Not All You Need," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2106.03253

[6] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/s10462-024-10884-2.

[7] A. Jha, R. Bhardwaj, P. Jha, and S. Jindal, "Empirical Comparison of Sleep Disorder Prediction Using Machine Learning Techniques," 2025, pp. 61–76. doi: 10.1007/978-981-97-8836-1_6.

[8] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[9] D. W. . Hosmer, S. Lemeshow, and R. X. . Sturdivant, *Applied logistic regression*. Wiley, 2013.

[10] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[11] J. H. Friedman, "REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1," 2001.

[12] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 1995. Available: http//roboticsStanfordedu/"ronnyk