Perbandingan *Random Forest* dan *XGBoost* Teroptimasi: Klasifikasi Stres berbasis Perilaku Digital dan Tidur

Comparing Optimized Random Forest and XGBoost: Stress Classification using Digital Behavior and Sleep

[1]M. Thoriqul Fadli, [2]M. Rudi Sanjaya*, [3]Dwi Rosa Indah, [4]M. Husni Syahbani
[1,2,3,4]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya
[1]Jl. Kutilang IV, Kec. Jakabaring, Kota Palembang, Sumatera Selatan, Indonesia
[2,3,4]Jl. Srijaya Negara, Bukit Lama, Kec. Ilir Bar. I, Kota Palembang, Sumatera Selatan, Indonesia
*e-mail: *09031282328037@student.unsri.ac.id*

***(received*:** 16 October 2026, ***revised*:** 29 October 2026, ***accepted*:** 5 November 2026)

Abstrak

Peningkatan durasi penggunaan perangkat digital berkaitan dengan berbagai risiko kesehatan mental, termasuk stres, sehingga diperlukan pendekatan prediktif yang mampu mengenali pola perilaku digital secara lebih objektif. Penelitian ini bertujuan membandingkan kinerja *Random Forest* dan XGBoost dalam klasifikasi tingkat stres berbasis perilaku digital dan tidur serta mengevaluasi pengaruh optimasi *hyperparameter* terhadap kinerja kedua model. Dataset Screentime vs Mental Wellness Survey 2025 yang terdiri atas 400 responden digunakan dengan distribusi 85,2% kelas "Stres" dan 14,8% kelas "Aman". Fitur yang terindikasi sebagai proksi target, yaitu Mental Wellness Index dan Productivity, dikeluarkan untuk mengurangi risiko *data leakage*. Penyeimbangan kelas dilakukan menggunakan SMOTENC di dalam pipeline pelatihan, sedangkan optimasi parameter dilakukan melalui GridSearchCV dengan *F1-Macro* sebagai metrik utama. Evaluasi utama menggunakan pembagian data latih dan uji sebesar 80:20 secara stratified pada 21 *random seed*, dengan *F1-Macro* dilaporkan sebagai rata-rata dari seluruh pembagian tersebut. Perbedaan kinerja dianalisis menggunakan *Wilcoxon signed-rank test* dan *5×2cv paired t-test*. Hasil menunjukkan bahwa rata-rata *F1-Macro Random Forest* meningkat dari 0,7721 menjadi 0,7807 setelah optimasi, sedangkan XGBoost meningkat dari 0,7370 menjadi 0,7684. *Random Forest* memperoleh *F1-Macro* lebih tinggi pada 11 dari 21 pembagian data. Uji *Wilcoxon* (p = 0,2959) dan *5×2cv paired t-test* (p = 0,5747) tidak menemukan perbedaan kinerja yang signifikan secara statistik. Analisis *feature importance* menunjukkan bahwa screen time hours dan sleep quality menjadi fitur dominan pada kedua model. Hasil penelitian ini terbatas pada dataset yang digunakan dan masih memerlukan validasi eksternal untuk menilai generalisasinya pada populasi lain.

Kata kunci: *digital behavior*, *hyperparameter optimization*, *Random Forest*, *stress classification*, XGBoost

Abstract

Increased digital device usage has been associated with various mental health risks, including stress, highlighting the need for predictive approaches capable of objectively identifying digital behavior patterns. This study aims to compare the performance of Random Forest and XGBoost in classifying stress levels based on digital behavior and sleep-related features, while also examining the effects of hyperparameter optimization on the performance of both models. The study uses the Screentime vs Mental Wellness Survey 2025 dataset, comprising 400 respondents, with a class distribution of 85.2% for the "Stressed" class and 14.8% for the "Safe" class. Features suspected of acting as proxies for the target variable, namely Mental Wellness Index and Productivity, were excluded to reduce the risk of data leakage. Class balancing was performed using SMOTENC within the training pipeline, while hyperparameter optimization was conducted using GridSearchCV, with F1-Macro as the primary optimization metric. The main evaluation employed stratified 80:20 training-test splits across 21 random seeds, with F1-Macro reported as the average across all splits. Differences in model performance were analyzed using the Wilcoxon signed-rank test and the 5×2 cross-validation paired t-test. The results show that the average F1-Macro score of Random Forest increased from 0.7721 to 0.7807 following optimization, while XGBoost improved from 0.7370 to 0.7684. Random Forest

 ISSN:2302-8149
e-ISSN:2540-9719

achieved a higher F1-Macro score in 11 out of 21 data splits. However, neither the Wilcoxon signed-rank test (p = 0.2959) nor the 5×2 cross-validation paired t-test (p = 0.5747) revealed a statistically significant difference in model performance. Feature importance analysis indicated that screen time hours and sleep quality were the dominant features in both models. These findings are limited to the dataset used in this study, and external validation is still required to assess the generalizability of the models to other populations.
Keywords: *digital behavior, hyperparameter optimization, Random Forest, stress classification, XGBoost*

1 Pendahuluan

Perangkat digital telah menjadi bagian dari aktivitas sehari-hari untuk bekerja, berkomunikasi, mengakses informasi, dan memperoleh hiburan. Penggunaan perangkat tersebut juga menjadi perhatian dalam kajian kesehatan karena durasi menatap layar (*screen time*) yang tinggi berasosiasi dengan berbagai keluhan fisik dan masalah kesehatan mental [1]. Pada populasi dewasa, tinjauan sistematis menunjukkan adanya asosiasi antara penggunaan layar dan berbagai aspek kesehatan mental, termasuk depresi, kecemasan, stres, dan kesejahteraan psikologis. Temuan tersebut bervariasi menurut jenis perangkat dan tujuan penggunaannya, sehingga hubungan perilaku digital dengan kesehatan mental perlu dipahami sesuai konteks aktivitas pengguna [2].

Bukti pada orang dewasa menunjukkan bahwa penggunaan ponsel pintar yang bermasalah berasosiasi dengan tingkat stres, kecemasan, dan depresi yang lebih tinggi serta kualitas tidur yang lebih buruk [3]. Namun, hubungan tersebut tidak seragam untuk seluruh aktivitas digital. Penelitian longitudinal yang menggabungkan data penelusuran internet dan kuesioner menemukan bahwa arah asosiasi penggunaan internet dengan stres berbeda menurut jenis aktivitas dan konteks penggunaannya [4]. Bukti intervensi turut memperlihatkan relevansi perilaku digital terhadap stres. Dalam uji acak terkontrol pada mahasiswa dewasa muda, kelompok yang menjalani pembatasan penggunaan ponsel selama tiga minggu menunjukkan tingkat stres dan gejala insomnia yang lebih rendah dibandingkan kelompok kontrol setelah intervensi [5].

Tidur merupakan aspek lain yang relevan dalam hubungan perilaku digital dengan stres. Penelitian pada populasi dewasa menemukan bahwa penggunaan layar sebelum tidur berasosiasi dengan durasi tidur yang lebih pendek dan kualitas tidur yang lebih buruk berdasarkan laporan responden [6]. Penelitian lain menunjukkan bahwa kualitas tidur yang buruk berasosiasi dengan tingkat stres yang lebih tinggi. Dalam penelitian tersebut, asosiasi penggunaan ponsel pada malam hari dengan stres terlihat pada laporan mandiri, tetapi tidak ditemukan secara jelas pada data pelacakan perangkat [7]. Pada mahasiswa dewasa muda, hubungan antara *screen time* dan stres juga bervariasi menurut durasi tidur dan tingkat aktivitas fisik [8]. Rangkaian temuan tersebut menjadi dasar bagi penelitian ini untuk menelaah fitur perilaku digital dan tidur secara bersama dalam klasifikasi stres.

Berdasarkan konteks tersebut, rumusan masalah penelitian ini mencakup tiga pertanyaan, yaitu bagaimana perbandingan kinerja *Random Forest* dan XGBoost dalam klasifikasi stres berbasis perilaku digital dan tidur, bagaimana perubahan kinerja kedua model setelah optimasi *hyperparameter*, serta fitur apa yang paling berkontribusi terhadap hasil klasifikasi. Sejalan dengan rumusan masalah tersebut, penelitian ini bertujuan membandingkan kinerja *Random Forest* dan XGBoost, mengevaluasi pengaruh optimasi *hyperparameter* terhadap kinerja kedua model, serta mengidentifikasi fitur yang berkontribusi terhadap klasifikasi stres.

Kontribusi penelitian ini berupa perbandingan empiris *Random Forest* dan XGBoost sebelum dan sesudah optimasi pada dataset Screentime vs Mental Wellness Survey 2025 menggunakan skema evaluasi yang sama. Ketidakseimbangan kelas ditangani menggunakan SMOTENC pada data latih, sedangkan optimasi *hyperparameter* dilakukan melalui GridSearchCV dengan *F1-Macro* sebagai metrik utama. Perbandingan disajikan berdasarkan rata-rata dan variasi kinerja pada 21 pembagian data secara *stratified*, disertai pengujian statistik untuk menelaah perbedaan kinerja kedua model. Analisis *feature importance* melengkapi perbandingan tersebut dengan mengidentifikasi kontribusi relatif fitur terhadap hasil klasifikasi, dengan interpretasi yang dibatasi pada model dan dataset yang dianalisis.

 ISSN:2302-8149
e-ISSN:2540-9719

2 Tinjauan Literatur

2.1 Perilaku Digital dan Tidur dalam Penelitian Stres

Penelitian mengenai perilaku digital dan stres pada populasi dewasa menunjukkan pentingnya mempertimbangkan konteks penggunaan perangkat serta perilaku keseharian pengguna. Belal dkk. menggabungkan rekaman aktivitas penelusuran internet dan pengukuran stres bulanan pada 1.490 pengguna internet di Jerman selama tujuh bulan. Melalui model *linear mixed-effects*, penelitian tersebut menemukan bahwa asosiasi penggunaan internet dengan stres berbeda menurut jenis aktivitas digital [4]. Sementara itu, Chandler dkk. menggunakan regresi hierarkis pada data potong lintang 513 mahasiswa dan melaporkan keterkaitan antara *screen time*, durasi tidur, serta interaksi keduanya dengan aktivitas fisik dalam menjelaskan variasi stres [8]. Kedua penelitian tersebut memberikan dasar untuk mempertimbangkan perilaku digital bersama aspek tidur, dengan tetap membatasi interpretasinya pada hubungan yang diamati.

Pemanfaatan data perilaku untuk pemodelan kesehatan mental juga dibahas oleh Khoo dkk. melalui tinjauan sistematis terhadap 184 penelitian. Kajian tersebut memetakan penggunaan pembelajaran mesin pada data multimodal yang diperoleh melalui *passive sensing*, termasuk media sosial, ponsel pintar, dan perangkat yang dikenakan pengguna [9]. Keragaman sumber data dan pendekatan pemodelan menunjukkan bahwa kinerja model perlu dinilai sesuai karakteristik data serta kondisi kesehatan mental yang menjadi sasaran. Dalam konteks penelitian ini, kajian asosiasi memberikan landasan bagi pemilihan kelompok fitur perilaku digital dan tidur, sedangkan kemampuan fitur tersebut untuk mengklasifikasikan stres perlu diperiksa melalui evaluasi prediktif pada data survei yang digunakan.

2.2 Klasifikasi Stres Menggunakan Machine Learning

Perbandingan algoritma pada penelitian terdahulu memperlihatkan hasil yang bergantung pada konfigurasi model dan karakteristik dataset. Fahria dkk. membandingkan *Random Forest* (RF), XGBoost, dan CatBoost untuk klasifikasi stres berdasarkan perilaku pengguna media sosial. RF memperoleh kinerja tertinggi pada pengujian tersebut, tetapi ketiga algoritma menggunakan parameter bawaan sehingga hasilnya terbatas pada konfigurasi yang diuji [10]. Pada konteks fitur fisiologis tidur, Tamboto dan Jaya menggunakan GridSearchCV untuk RF dan XGBoost. Kedua model memperoleh akurasi pengujian yang sama setelah pembulatan, sedangkan RF menghasilkan rata-rata *cross-validation* yang lebih tinggi [11]. Perbedaan konteks kedua penelitian tersebut menunjukkan perlunya membedakan hasil perbandingan parameter bawaan dari hasil perbandingan setelah optimasi.

Pengaruh optimasi juga perlu ditelaah bersama tahapan pengolahan data yang menyertainya. Sugiharto dan Nataliani membandingkan Logistic Regression, K-Nearest Neighbors, dan Support Vector Machine melalui skenario baseline, SMOTE, seleksi fitur menggunakan *Recursive Feature Elimination* (RFE), serta optimasi *hyperparameter*. SVM memperoleh hasil tertinggi pada skenario gabungan, sehingga pencapaiannya perlu dibaca dalam konteks keseluruhan tahapan tersebut [12]. Pada data EEG, Sifaunnufus Ms dkk. justru melaporkan bahwa RF dengan fitur *Power Spectral Density* (PSD), tanpa SMOTE dan seleksi fitur menggunakan *Particle Swarm Optimization* (PSO), memberikan hasil terbaik di antara konfigurasi yang diuji [13]. Temuan ini menunjukkan bahwa penambahan tahapan penyeimbangan atau seleksi fitur perlu dievaluasi secara empiris sesuai model dan dataset.

Penelitian dengan sasaran yang berdekatan turut memberikan pembanding metodologis. Hafin dkk. menggunakan RF dan Logistic Regression untuk klasifikasi risiko depresi pada remaja berdasarkan perilaku media sosial dan indikator kesehatan mental [14]. Adapun Pratama Putra dkk. menggunakan dataset *Screentime vs Mental Wellness Survey 2025* untuk membandingkan RF, SVR, dan XGBoost dalam prediksi skor stres kontinu [15]. Relevansi kedua penelitian tersebut terletak pada pemanfaatan fitur perilaku dan pendekatan pemodelannya, dengan perbedaan sasaran prediksi yang perlu diperhatikan. Ringkasan dataset, metode, evaluasi, hasil, serta keterbatasan atau relevansi penelitian terdahulu disajikan pada Tabel 1.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 Ringkasan penelitian terdahulu

Penulis dan tahun	*Dataset*	Metode	Skema evaluasi	Hasil utama	Keterbatasan atau relevansi
Fahria dkk. (2026) [10]	*Mental Health and Social Media Balance Dataset*; 667 data; klasifikasi stres biner.	RF, XGBoost, dan CatBoost dengan parameter bawaan.	Pembagian latih–uji 80:20 secara *stratified*.	RF: akurasi dan *F1-Score* 0,84. XGBoost: akurasi 0,78 dan *F1-Score* 0,80.	Parameter belum dioptimasi; distribusi kelas relatif seimbang; validasi eksternal belum dilakukan.
Tamboto dan Jaya (2026) [11]	*Human Stress Detection in and through Sleep*; 630 data; lima kelas stres.	RF dan XGBoost dengan GridSearchCV.	Pembagian 80:20 dan *cross-validation*; jumlah fold tidak dirinci.	Akurasi uji kedua model 0,980. Rata-rata CV: RF 0,99206 dan XGBoost 0,98730.	Menggunakan fitur fisiologis tidur. Ukuran variasi hasil CV tidak disajikan sehingga rata-rata yang lebih tinggi belum menunjukkan kestabilan yang lebih baik.
Sifaunnufus Ms dkk. (2026) [13]	SAM40; 40 subjek dan 360 sampel EEG; klasifikasi stres biner.	RF, XGBoost, SVM, Logistic Regression, dan *stacking*; variasi fitur PSD, Hjorth, Higuchi, SMOTE, serta seleksi fitur PSO.	306 sampel latih dan 54 sampel uji.	RF dengan PSD tanpa SMOTE dan PSO: akurasi 0,7778 dan *F1-Score* 0,7707.	Modalitas data berbeda. Hasil menunjukkan bahwa manfaat penyeimbangan dan seleksi fitur bergantung pada konfigurasi yang diuji.
Hafin dkk. (2026) [14]	*Teenager Mental Health Dataset*; 1.200 data; klasifikasi risiko depresi.	RF, Logistic Regression, dan SMOTE.	Pembagian data latih–uji; rasio pembagian tidak dirinci.	RF: akurasi dan *F1-Score* 0,9979; ROC-AUC 1,0000.	Target berupa depresi. SMOTE dilaporkan dilakukan sebelum pemisahan data; penulis menyarankan validasi tambahan untuk menilai generalisasi.
Sugiharto dan Nataliani (2026) [12]	*Student Lifestyle Dataset*; 2.000 data; tiga kelas stres.	Logistic Regression, KNN, dan SVM; skenario SMOTE, RFE, serta GridSearchCV.	Pembagian 80:20 dengan seed 42; tuning menggunakan 5-fold StratifiedK Fold.	SVM pada skenario gabungan: akurasi 0,98 dan *F1-Score* 0,97.	Algoritma pembanding berbeda; hasil skenario akhir mencakup gabungan penyeimbangan, seleksi fitur, dan tuning.
Pratama Putra dkk. (2026) [15]	*Screentime vs Mental Wellness Survey 2025*; 400 observasi; skor stres kontinu.	RF Regression, SVR, dan XGBoost Regression; penyetelan parameter dan analisis SHAP.	Pembagian 80:20; 10-fold CV pada data latih.	XGBoost: MAE 0,6502; RMSE 0,8253; R^2 0,8367.	Menggunakan dataset yang sama, tetapi target berupa regresi serta tetap menyertakan Mental Wellness Index dan Productivity sebagai prediktor.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 memperlihatkan bahwa perbedaan sasaran prediksi, kelompok fitur, dan skema evaluasi membatasi perbandingan langsung antarangka kinerja. Penelitian dengan akurasi lebih tinggi belum tentu menyediakan model yang lebih sesuai untuk dataset atau sasaran penelitian lain. Dalam kumpulan studi yang ditinjau, perbandingan RF dan XGBoost masih perlu diperjelas untuk skenario klasifikasi stres biner dengan himpunan fitur yang dibatasi, penilaian sebelum dan sesudah optimasi, serta pemeriksaan variasi hasil pada beberapa pembagian data.

2.3 Penyeimbangan Kelas dan Keandalan Evaluasi Model

Ketidakseimbangan kelas merupakan salah satu pertimbangan dalam mengevaluasi model klasifikasi stres. Altalhan dkk. menjelaskan bahwa distribusi kelas yang tidak seimbang dapat menyebabkan model lebih mengutamakan kelas mayoritas, sehingga evaluasi perlu dilengkapi metrik seperti recall, *F1-Score*, dan AUC [16]. Penggunaan SMOTE dapat membantu meningkatkan representasi kelas minoritas, tetapi kualitas sampel sintetis tetap bergantung pada karakteristik data. Matharaarachchi dkk. menunjukkan bahwa observasi abnormal pada kelas minoritas dapat menghasilkan sampel sintetis yang kurang representatif dan mengusulkan pengembangan SMOTE untuk menangani kondisi tersebut [17]. Dengan demikian, penggunaan penyeimbangan kelas perlu disertai pemeriksaan kinerja pada masing-masing kelas.

Keandalan evaluasi juga berkaitan dengan variasi hasil yang diperoleh selama pengembangan model. Huang dan Huang membandingkan pendekatan analitis dengan simulasi bootstrap untuk mengkarakterisasi distribusi metrik evaluasi, sehingga kajiannya menekankan pentingnya mempertimbangkan ketidakpastian di sekitar nilai kinerja [18]. Riley dan Collins membahas kestabilan prediksi sebagai hasil interaksi antara ukuran data, kompleksitas model, dan strategi pengembangan. Pembatasan kompleksitas dapat membantu pada kondisi tertentu, sedangkan proses pemilihan *hyperparameter* berbasis data juga dapat menambah ketidakpastian [19]. Kedua kajian tersebut memberikan alasan untuk memperhatikan variasi hasil selain nilai kinerja tunggal, tanpa menganggap bahwa optimasi selalu menghasilkan model yang lebih stabil.

Urutan pengolahan data turut menentukan interpretasi hasil evaluasi. Apicella dkk. menjelaskan risiko *synthesis leakage* ketika sampel sintetis dibentuk sebelum pemisahan data latih dan evaluasi [20]. Hal ini relevan dengan prosedur yang dilaporkan Hafin dkk., yang menempatkan SMOTE sebelum pembagian data [14]. Urutan tersebut perlu dipertimbangkan ketika menafsirkan hasilnya karena dapat membawa informasi dari calon data uji ke proses pembentukan sampel sintetis [20]. Penerapan SMOTE pada bagian pelatihan di setiap fold merupakan langkah untuk mengurangi risiko kebocoran pada tahap penyeimbangan, sedangkan tahapan prapemrosesan lainnya tetap memerlukan pemeriksaan tersendiri.

Berdasarkan kajian tersebut, penelitian ini berfokus pada perbandingan RF dan XGBoost untuk klasifikasi stres biner menggunakan fitur demografis, perilaku digital, dan tidur, tanpa menyertakan Mental Wellness Index dan Productivity sebagai prediktor. Pengaruh optimasi *hyperparameter* ditelaah melalui perbandingan kondisi baseline dan teroptimasi yang sama-sama menggunakan SMOTENC. Evaluasi pada 21 pembagian data digunakan untuk menilai rata-rata dan variasi kinerja, dengan perhatian pada *F1-Macro* serta pengenalan kelas minoritas. Analisis *feature importance* melengkapi perbandingan tersebut dengan mengidentifikasi kontribusi fitur terhadap prediksi model. Dengan rancangan ini, kontribusi penelitian diarahkan pada bukti komparatif dalam skenario pemodelan yang ditetapkan, dengan mempertimbangkan perbedaan konfigurasi dan keterbatasan evaluasinya.

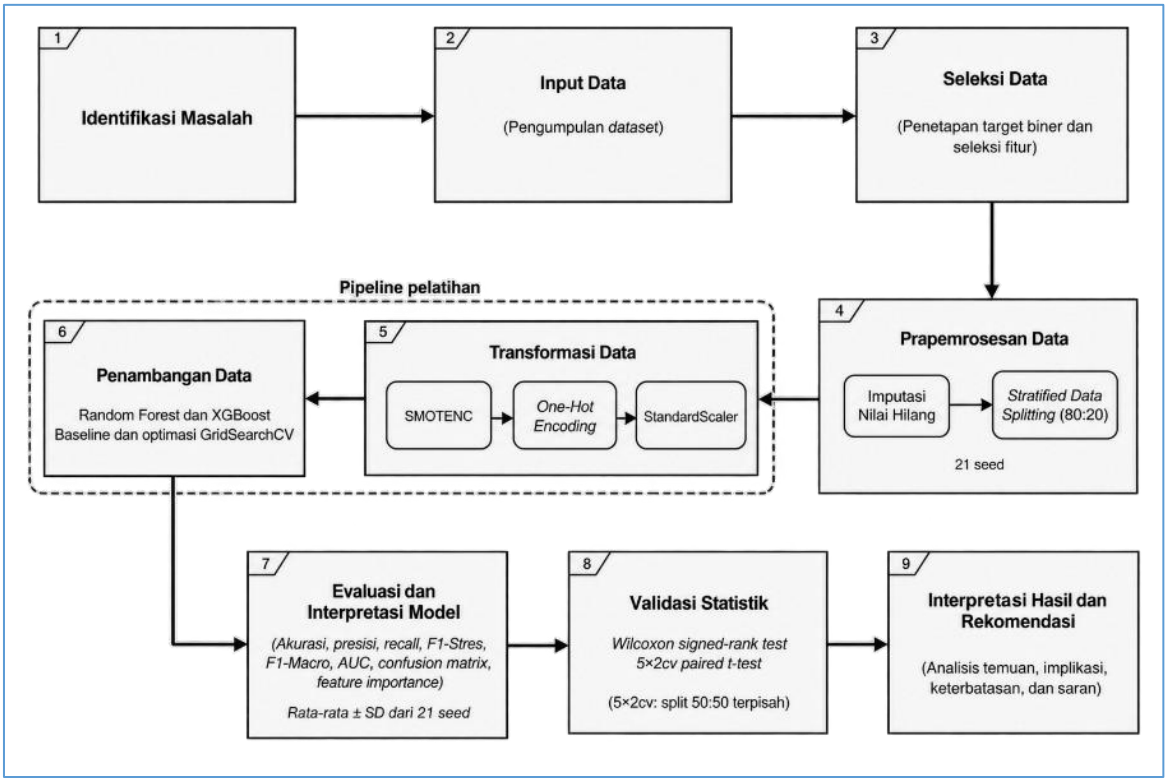
3 Metode Penelitian

3.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan eksperimental kuantitatif untuk membandingkan *Random Forest* dan XGBoost dalam klasifikasi stres berbasis perilaku digital dan tidur, mengevaluasi pengaruh optimasi *hyperparameter*, serta mengidentifikasi fitur yang berkontribusi terhadap prediksi. Tahapan penelitian meliputi pengumpulan data, pembentukan target biner, seleksi fitur, prapemrosesan, pelatihan model *baseline* dan teroptimasi, serta evaluasi dan interpretasi hasil. Evaluasi utama menggunakan pembagian data 80:20 secara *stratified* pada 21 *random seed*,

 ISSN:2302-8149
e-ISSN:2540-9719

sedangkan pengujian 5×2cv menggunakan pembagian 50:50 secara terpisah. Alur penelitian disajikan pada Gambar 1.



Gambar 1 Alur penelitian

3.2 Pengumpulan dan Persiapan *Dataset*

Data sekunder diperoleh dari dataset publik *Screentime vs Mental Wellness Survey 2025* yang diunggah oleh Adharshini Kumaresan melalui Kaggle. Dataset memuat 400 rekaman dengan atribut demografis, perilaku digital, tidur, aktivitas fisik dan sosial, serta skor stres dan kesejahteraan mental. Usia responden berkisar 16–60 tahun, dengan 382 responden (95,5%) berusia ≥18 tahun. Dengan demikian, dataset didominasi orang dewasa, tetapi tidak terbatas pada kelompok tersebut.

Variabel *stress_level_0_10* dikonversi menjadi target biner: skor ≥6 diberi label "Stres" (1), sedangkan skor <6 diberi label "Aman" (0). Ambang 6 dipilih berdasarkan eksplorasi awal pemodelan dan dipertahankan pada seluruh eksperimen utama sebagai batas operasional, bukan batas diagnosis klinis. Pembentukan target menghasilkan 341 sampel kelas "Stres" (85,2%) dan 59 sampel kelas "Aman" (14,8%). Karakteristik dataset dirangkum pada Tabel 2.

Tabel 2 Karakteristik distribusi *dataset*

Karakteristik	Nilai
Jumlah *Record*	400
Atribut awal	15 atribut substantif dan 1 kolom kosong
Rentang usia	16–60 tahun
Responden berusia ≥18 tahun	382 (95,5%)
Kelas "Stres" / "Aman"	341 (85,2%) / 59 (14,8%)
Nilai hilang pada 15 atribut substantif	Tidak ditemukan
Prediktor setelah seleksi	11 atribut: 8 numerik dan 3 kategorikal

 ISSN:2302-8149
e-ISSN:2540-9719

3.3 Seleksi Fitur dan Mitigasi Kebocoran Data (*Data Leakage*)

Seleksi fitur dilakukan untuk membatasi prediktor pada karakteristik demografis dan perilaku serta mengurangi risiko penggunaan informasi yang merepresentasikan target. Kebocoran data dapat menghasilkan evaluasi kinerja yang terlalu optimistis [21], [22]. Atribut *User ID* dikeluarkan karena merupakan pengenal responden, sedangkan *Stress Level* dikeluarkan karena digunakan untuk membentuk target klasifikasi.

Mental Wellness Index dan *Productivity* juga dikeluarkan sebagai fitur yang dicurigai menjadi proksi target. Pemeriksaan awal menunjukkan korelasi Pearson dengan skor stres masing-masing sebesar −0,914 dan −0,880. Model *decision stump* menggunakan *Mental Wellness Index* sebagai satu-satunya prediktor mencapai akurasi 92,5% pada pembagian 80:20 dengan *seed* 42. Temuan tersebut menjadi pertimbangan untuk membatasi model pada fitur demografis dan perilaku, tetapi tidak membuktikan bahwa kedua atribut tersebut merupakan turunan matematis target. Keputusan seleksi fitur disajikan pada Tabel 3.

Tabel 3 Matriks keputusan seleksi fitur

Fitur	Status	Alasan
User ID	Dieksklusi	Pengenal responden
Stress Level	Dieksklusi	Sumber pembentukan target biner
Mental Wellness Index	Dieksklusi	Dugaan proksi target; korelasi dengan skor stres −0,914
Productivity	Dieksklusi	Dugaan proksi target; korelasi dengan skor stres −0,880
Age, Gender, Occupation, Work Mode	Dipertahankan	Fitur demografis (karakteristik responden)
Screen Time Hours, Work Screen Hours, Leisure Screen Hours, Sleep Hours, Sleep Quality, Exercise Minutes per Week, Social Hours per Week	Dipertahankan	Fitur perilaku digital, tidur, serta aktivitas fisik dan sosial.

3.4 Pra-pemrosesan dan *Pipeline* Penyeimbangan Data Terisolasi

Nama kolom diseragamkan, dan kolom kosong *unnamed: 15* dikeluarkan. Prosedur imputasi menggunakan median untuk atribut numerik dan modus untuk atribut kategorikal dijalankan sebelum pembagian data, di luar *pipeline*. Pada dataset yang digunakan, tidak ditemukan nilai hilang pada atribut yang dipertahankan sehingga prosedur imputasi tidak mengubah nilai prediktor. Data kemudian dibagi secara *stratified* menjadi 320 sampel latih dan 80 sampel uji. Data latih terdiri atas 273 sampel “Stres” dan 47 sampel “Aman”, sedangkan data uji terdiri atas 68 sampel “Stres” dan 12 sampel “Aman”.

Penyeimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique for Nominal and Continuous* (SMOTENC) karena prediktor mencakup atribut numerik dan kategorikal. *Gender*, *Occupation*, dan *Work Mode* ditetapkan sebagai atribut kategorikal dan diberikan kepada SMOTENC sebelum *one-hot encoding*. Pengaturan menggunakan lima tetangga terdekat dan menambah sampel kelas minoritas hingga jumlahnya sama dengan kelas mayoritas pada data pelatihan. Pembentukan nilai sintetis pada komponen numerik mengikuti interpolasi pada Persamaan (1), sedangkan nilai kategorikal dipilih berdasarkan kategori yang paling sering muncul di antara tetangga kelas minoritas.

$$x_{new} = x_i + \lambda \cdot (x_{nn} - x_i) \quad (1)$$

 ISSN:2302-8149
e-ISSN:2540-9719

Pada persamaan (1), x_{new} adalah nilai numerik sintetis, x_i adalah nilai pada sampel minoritas acuan, x_{nn} adalah nilai pada salah satu tetangganya, dan λ merupakan bilangan acak pada rentang 0–1. Interpolasi ini berlaku pada komponen numerik; kategori tidak diinterpolasi menjadi nilai pecahan.

Urutan *pipeline* pelatihan adalah SMOTENC, *One-Hot Encoding* dengan pengaturan *drop first*, StandardScaler, dan model klasifikasi. Pada optimasi, SMOTENC hanya diterapkan pada bagian latih setiap *fold*, sedangkan encoder dan StandardScaler dipelajari dari data pelatihan hasil penyeimbangan. Data validasi dan uji ditransformasikan menggunakan encoder dan scaler yang telah dipelajari tanpa penambahan sampel sintetis. Dengan demikian, sebagaimana ditunjukkan pada Gambar 1, imputasi berada di luar *pipeline*, sedangkan encoding berada di dalam *pipeline* setelah SMOTENC.

3.5 Algoritma Klasifikasi

Penelitian ini membandingkan dua algoritma *ensemble* berbasis pohon. *Random Forest* membangun sejumlah pohon keputusan menggunakan sampel *bootstrap* dan pemilihan sebagian fitur secara acak pada pemisahan simpul, kemudian menggabungkan prediksi seluruh pohon [23]. XGBoost menggunakan *gradient boosting*, yaitu penambahan pohon secara berurutan untuk memperbaiki fungsi objektif, dengan regularisasi untuk mengendalikan kompleksitas model [24]. Perbedaan mekanisme pembelajaran tersebut menjadi dasar perbandingan kedua algoritma pada fitur dan skema evaluasi yang sama.

3.6 Optimasi *Hyperparameter*

Model *baseline* didefinisikan sebagai *Random Forest* dan XGBoost dengan parameter bawaan pustaka, selain pengaturan *random_state* pada kedua algoritma serta *eval_metric='logloss'* pada XGBoost. Kedua *baseline* menggunakan *pipeline* SMOTENC, encoding, dan standardisasi yang sama dengan model teroptimasi. Perbandingan sebelum dan sesudah optimasi dilakukan pada pembagian data yang sama agar perubahan kinerja dapat dinilai dalam rancangan evaluasi yang setara.

Optimasi menggunakan GridSearchCV dengan *3-fold stratified cross-validation* pada data latih. *F1-Macro* digunakan sebagai fungsi penilaian karena memberikan bobot yang sama pada kedua kelas. Kombinasi parameter dipilih berdasarkan rata-rata *F1-Macro* tertinggi pada validasi internal, kemudian *pipeline* dilatih kembali pada seluruh data latih dan dievaluasi pada data uji. Proses ini diulang pada setiap pembagian data. Ruang pencarian mencakup 12 kombinasi untuk *Random Forest* dan 27 kombinasi untuk XGBoost, sebagaimana disajikan pada Tabel 4.

Tabel 4 Ruang pencarian (*search space*) *hyperparameter model*

Algoritma	*Hyperparameter*	Ruang Nilai pengujian	Total Kombinasi
Random Forest	n_estimators	100, 200	12
	max_depth	None, 5, 10	
	min_samples_split	2, 5	
XGBoost	n_estimators	50, 100, 200	27
	learning_rate	0,01; 0,05; 0,1	
	max_depth	3,5,7	

3.7 Metrik Evaluasi Kinerja

Evaluasi menggunakan akurasi, presisi dan *recall* kelas "Stres", *F1-Stres*, *recall* kelas "Aman", *F1-Macro*, serta AUC. *F1-Macro* merupakan metrik utama, sedangkan *recall* kelas "Aman" menunjukkan kemampuan mengenali kelas minoritas. AUC dihitung dari probabilitas prediksi kelas "Stres", bukan label prediksi biner. Ringkasan utama melaporkan rata-rata dan simpangan baku akurasi, *F1-Macro*, *recall* kelas "Aman", serta AUC dari 21 pembagian data.

Sebagai pembanding sederhana, prediksi yang selalu memilih kelas mayoritas "Stres" menghasilkan akurasi 68/80 = 85,0% pada data uji. Pembanding ini berbeda dari *baseline* algoritma yang dijelaskan pada Subbab 3.6. *Confusion matrix*, metrik per kelas, dan *feature importance* dari model teroptimasi pada *seed* 42 disajikan sebagai ilustrasi. *Feature importance* diperoleh dari atribut

 ISSN:2302-8149
e-ISSN:2540-9719

*feature_importances_* masing-masing model dan digunakan untuk menjelaskan kontribusi relatif fitur terhadap prediksi, tanpa menyimpulkan hubungan sebab-akibat.

3.8 Uji *Robustness* dan Signifikansi Statistik

Evaluasi diulang pada 21 *random seed* dengan proporsi data latih dan uji 80:20 serta prosedur pelatihan yang sama. Kedua algoritma menggunakan pasangan data latih dan uji yang identik pada setiap *seed*. Perbedaan *F1-Macro* model teroptimasi dianalisis menggunakan *Wilcoxon signed-rank test* dua sisi dengan taraf signifikansi 0,05. Karena pembagian dilakukan berulang pada dataset yang sama, sebagian observasi dapat muncul kembali pada beberapa eksperimen sehingga hasil antarpembagian tidak sepenuhnya independen. Oleh karena itu, nilai p Wilcoxon diperlakukan sebagai bukti indikatif, sedangkan simpangan baku menggambarkan variasi kinerja pada pembagian dataset tersebut.

Pengujian tambahan menggunakan *5×2cv paired t-test* yang diperkenalkan oleh Dietterich [25]. Prosedur ini mencakup lima replikasi pembagian data 50:50 secara *stratified*. Pada setiap replikasi, peran kedua bagian sebagai data latih dan uji dipertukarkan sehingga diperoleh sepuluh pengukuran perbedaan *F1-Macro*. Optimasi GridSearchCV dengan validasi internal tiga *fold* dilakukan kembali pada data latih di setiap arah pengujian. Pengujian ini dijalankan terpisah dari evaluasi utama 21 *seed*, dengan taraf signifikansi 0,05. Hasil pengujian ditafsirkan bersama besarnya selisih dan variasi kinerja; hasil yang tidak signifikan tidak digunakan untuk menyatakan kedua algoritma setara.

4 Hasil dan Pembahasan

Bagian ini menyajikan hasil eksperimen klasifikasi tingkat stres menggunakan *Random Forest* dan XGBoost setelah melalui tahapan mitigasi kebocoran data, penyeimbangan kelas menggunakan SMOTENC, serta optimasi *hyperparameter* dengan GridSearchCV. Analisis hasil diawali dengan validasi diagnostik terhadap fitur yang berpotensi menjadi *proxy* target untuk memastikan bahwa model dibangun berdasarkan variabel demografis dan perilaku digital yang relevan. Selanjutnya, kinerja kedua algoritma dievaluasi melalui perbandingan model baseline dan model hasil optimasi menggunakan metrik *Accuracy*, *F1-Score*, *Recall* kelas minoritas, *F1-Macro*, dan AUC. Hasil prediksi kemudian dianalisis lebih lanjut melalui *Confusion Matrix* dan *feature importance* untuk mengidentifikasi pola kesalahan klasifikasi serta kontribusi masing-masing fitur terhadap keputusan model. Ketahanan model dievaluasi pada 21 pembagian data secara *stratified* (80:20), dengan rata-rata dan simpangan baku kinerja sebagai hasil utama, sedangkan hasil *seed* 42 disajikan sebagai ilustrasi. Perbedaan kinerja kedua model selanjutnya dianalisis menggunakan pengujian statistik.

4.1 Hasil Diagnostik Fitur dan Validasi Dataset Akhir

Hasil diagnostik menunjukkan bahwa *Mental Wellness Index* dan *Productivity* memiliki korelasi Pearson dengan skor stres kontinu masing-masing sebesar −0,914 dan −0,880. Korelasi *Mental Wellness Index* dengan target biner sebesar −0,745. Sebagai pembanding, korelasi *Screen Time Hours* dengan skor stres sebesar 0,698, sedangkan *Sleep Quality* sebesar −0,559. Rincian korelasi fitur numerik dengan skor stres disajikan pada Tabel 5.

Tabel 5 Hasil korelasi fitur numerik terhadap tingkat stres

Fitur	Korelasi terhadap Stress Level	Status
Mental Wellness Index	-0.914	Dieksklusi
Productivity	-0.880	Dieksklusi
Screen Time Hours	0.698	Dipertahankan
Sleep Quality	-0.559	Dipertahankan
Leisure Screen Hours	0.515	Dipertahankan
Sleep Hours	-0.481	Dipertahankan
Work Screen Hours	0.307	Dipertahankan

 ISSN:2302-8149
e-ISSN:2540-9719

Fitur	Korelasi terhadap Stress Level	Status
Exercise Minutes per Week	-0.160	Dipertahankan
Social Hours per Week	-0.114	Dipertahankan
Age	0.046	Dipertahankan

Pemeriksaan menggunakan *decision stump* dengan kedalaman satu dan *Mental Wellness Index* sebagai satu-satunya prediktor menghasilkan akurasi 92,50% pada data uji *seed* 42. Pada 320 sampel data latih dari pembagian yang sama, korelasi *Mental Wellness Index* dan *Productivity* dengan skor stres masing-masing sebesar −0,909 dan −0,871. Hubungan yang kuat tersebut menjadi pertimbangan untuk mengeluarkan kedua atribut sebagai fitur yang dicurigai menjadi proksi target, tetapi tidak membuktikan bahwa keduanya merupakan turunan matematis target.

Dataset akhir mempertahankan 11 prediktor, terdiri atas delapan atribut numerik dan tiga atribut kategorikal. Prediktor tersebut mencakup karakteristik demografis dan pekerjaan, durasi penggunaan layar, pola tidur, serta aktivitas fisik dan sosial. Pemodelan selanjutnya membandingkan kedua algoritma pada kumpulan fitur yang sama.

4.2 Hasil Optimasi *Hyperparameter* dan Kinerja Model

Hasil utama diperoleh dari 21 pembagian data yang sama untuk kedua algoritma. Pada setiap pembagian, model *baseline* dan teroptimasi menggunakan *pipeline* SMOTENC yang sama; optimasi dilakukan melalui validasi internal pada data latih. Tabel 6 menyajikan rata-rata ± simpangan baku akurasi, *F1-Macro*, *recall* kelas "Aman", dan AUC, beserta pembanding prediksi kelas mayoritas.

Tabel 6 Kinerja agregat model dan pembanding mayoritas pada 21 *seed*

Algoritma	Kondisi	Akurasi	*F1-Macro*	*Recall* Aman	AUC
Random Forest	*Baseline*	0,8774 ± 0,0341	0,7721 ± 0,0566	0,6587 ± 0,1205	0,9218 ± 0,0329
Random Forest	Teroptimasi	0,8744 ± 0,0361	0,7807 ± 0,0595	0,7341 ± 0,1281	0,9257 ± 0,0320
XGBoost	*Baseline*	0,8637 ± 0,0298	0,7370 ± 0,0560	0,5754 ± 0,1417	0,9097 ± 0,0280
XGBoost	Teroptimasi	0,8673 ± 0,0300	0,7684 ± 0,0431	0,7143 ± 0,1195	0,9145 ± 0,0297
Prediksi Mayoritas	Selalu "Stres"	0,8500 ± 0,0000	—	—	—

Catatan: nilai model disajikan sebagai rata-rata ± simpangan baku. Pembanding mayoritas dicantumkan untuk akurasi; tanda "—" berarti metrik tidak disajikan pada baris tersebut.

Distribusi kelas keseluruhan sebesar 85,2% "Stres" dan 14,8% "Aman" memberikan perkiraan akurasi mayoritas sekitar 85%. Pada setiap data uji terdapat 68 sampel "Stres" dari 80 sampel, sehingga prediksi yang selalu memilih "Stres" menghasilkan akurasi tepat 85,0%. Pembanding ini berbeda dari *baseline* algoritma sebelum optimasi. Rata-rata akurasi keempat konfigurasi model berada di atas pembanding mayoritas tersebut.

Optimasi meningkatkan rata-rata *F1-Macro Random Forest* dari 0,7721 menjadi 0,7807 dan *recall* kelas "Aman" dari 0,6587 menjadi 0,7341, sementara akurasi sedikit menurun dari 0,8774 menjadi 0,8744. Pada XGBoost, rata-rata *F1-Macro* meningkat dari 0,7370 menjadi 0,7684, *recall* kelas "Aman" dari 0,5754 menjadi 0,7143, dan akurasi dari 0,8637 menjadi 0,8673. Rata-rata AUC juga meningkat pada kedua algoritma. Hasil ini menunjukkan bahwa perubahan kinerja setelah optimasi perlu dibaca per metrik, sesuai penggunaan *F1-Macro* sebagai dasar pemilihan parameter.

 ISSN:2302-8149
e-ISSN:2540-9719

Pada kondisi teroptimasi, *Random Forest* memperoleh rata-rata *F1-Macro* 0,7807 dan AUC 0,9257, sedangkan XGBoost memperoleh 0,7684 dan 0,9145. Perbedaan ini merupakan perbandingan deskriptif; variasi antarpembagian dan hasil pengujian statistik dibahas pada Subbab 4.5 dan 4.6.

Untuk melengkapi hasil agregat, Tabel 7 menunjukkan konfigurasi terpilih pada *seed* 42 sebagai ilustrasi. Pada pembagian ini, *Random Forest* menggunakan *min_samples_split* = 5. Konfigurasi tersebut merupakan hasil validasi internal pada pembagian yang bersangkutan dan dapat berbeda pada *seed* lain.

Tabel 7 *Hyperparameter* terpilih pada seed 42 sebagai ilustrasi

Algoritma	*Hyperparameter*	Nilai Terbaik
Random Forest	n_estimators	100
	max_depth	5
	min_samples_split	5
XGBoost	n_estimators	100
	learning_rate	0.1
	max_depth	3

Rincian kinerja pada *seed* 42 disajikan pada Tabel 8. *Random Forest* teroptimasi memperoleh akurasi 0,9250 dan *F1-Macro* 0,8529, sedangkan XGBoost teroptimasi memperoleh akurasi 0,9000 dan *F1-Macro* 0,7714. Nilai tersebut menggambarkan satu pembagian data dan tidak menggantikan hasil agregat pada Tabel 6.

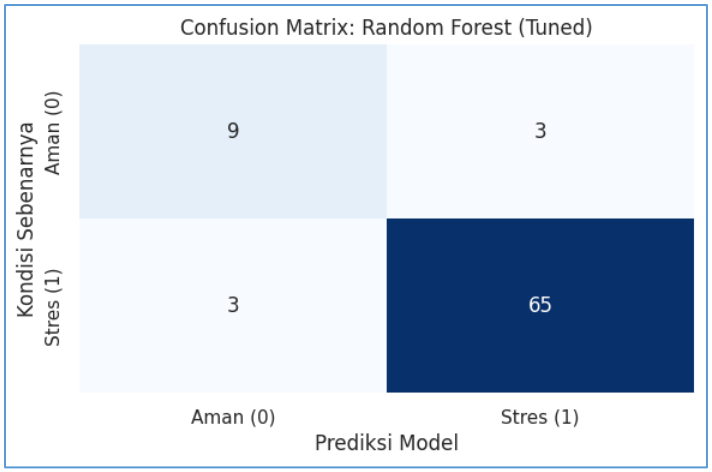
Tabel 8 Kinerja pada seed 42 sebagai ilustrasi

Metrik	RF *Baseline*	RF teroptimasi	XGBoost *Baseline*	XGBoost teroptimasi
Akurasi	0,8875	0,9250	0,8625	0,9000
Presisi Stres	0,9041	0,9559	0,8904	0,9167
Recall Stres	0,9706	0,9559	0,9559	0,9706
F1-Stres	0,9362	0,9559	0,9220	0,9429
Recall Aman	0,4167	0,7500	0,3333	0,5000
F1-Macro	0,7312	0,8529	0,6715	0,7714
AUC	0,9602	0,9706	0,9449	0,9510

4.3 Analisis Kesalahan Klasifikasi Model

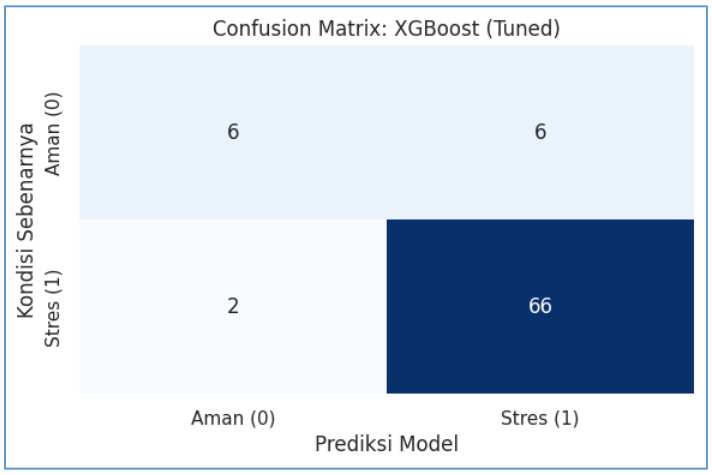
Pola kesalahan model teroptimasi diilustrasikan melalui *confusion matrix* pada *seed* 42. Gambar 2 menunjukkan bahwa *Random Forest* mengklasifikasikan dengan benar sembilan dari 12 sampel “Aman” dan 65 dari 68 sampel “Stres”. Tiga sampel “Aman” diprediksi sebagai “Stres”, sedangkan tiga sampel “Stres” diprediksi sebagai “Aman”. Dengan demikian, terdapat 74 prediksi benar dari 80 sampel, sesuai akurasi 0,9250 pada Tabel 8.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 2 *Confusion matrix random forest* teroptimasi pada *seed* 42

Pada Gambar 3, XGBoost mengenali enam dari 12 sampel "Aman" dan 66 dari 68 sampel "Stres", menghasilkan 72 prediksi benar dari 80 sampel. Pada pembagian ini, *recall* kelas "Aman" *Random Forest* lebih tinggi, yaitu 0,7500 dibandingkan 0,5000, sedangkan *recall* kelas "Stres" XGBoost lebih tinggi, yaitu 0,9706 dibandingkan 0,9559. Perbedaan pola kesalahan tersebut menjelaskan nilai *F1-Macro* pada ilustrasi ini; perbandingan umum kedua model tetap mengacu pada agregat 21 *seed*.

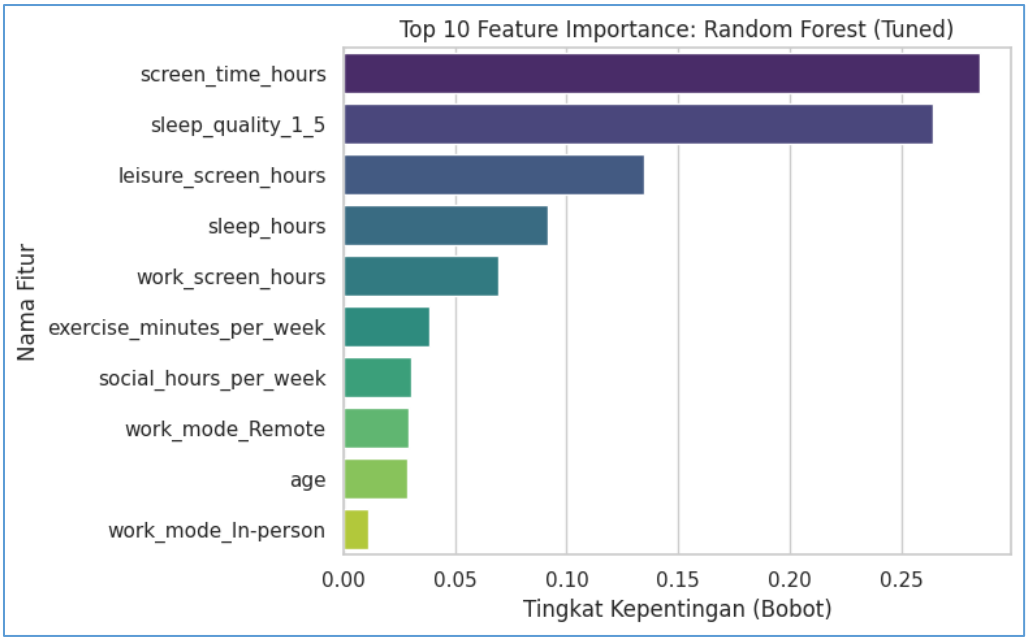


Gambar 3 *Confusion matrix* XGBoost teroptimasi pada *seed* 42

4.4 Analisis Kontribusi Fitur terhadap Prediksi Stres

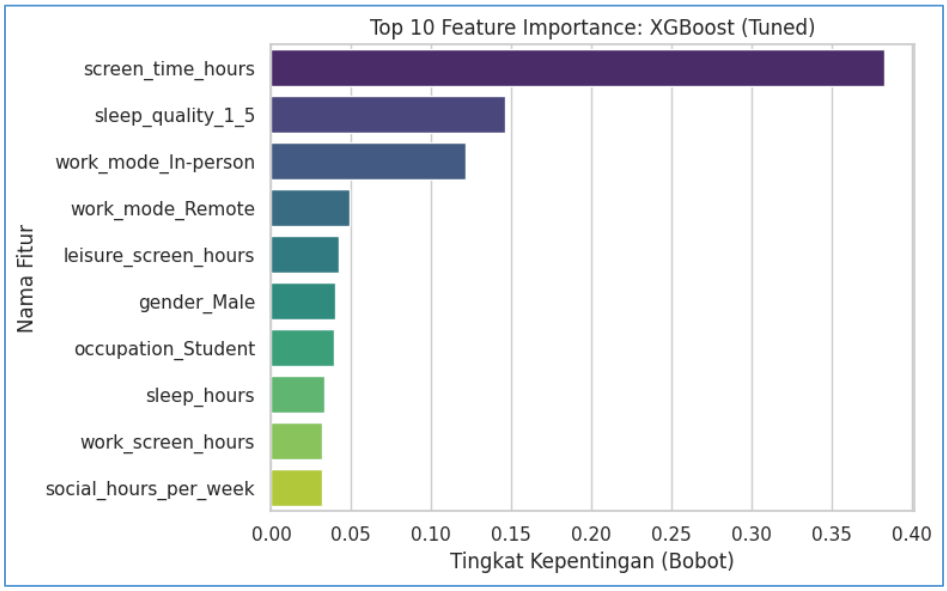
Analisis *feature importance* menggunakan model teroptimasi pada *seed* 42. Pada *Random Forest*, *Screen Time Hours* dan *Sleep Quality* menempati dua urutan teratas, diikuti *Leisure Screen Hours*, *Sleep Hours*, dan *Work Screen Hours*. Aktivitas fisik dan sosial, kategori mode kerja, serta usia memiliki bobot lebih kecil dalam sepuluh fitur yang ditampilkan pada Gambar 4.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 4 Sepuluh fitur terpenting *Random Forest* teroptimasi pada *seed* 42

Pada XGBoost, dua fitur teratas juga berupa *Screen Time Hours* dan *Sleep Quality*. Urutan berikutnya adalah *work_mode_In-person*, *work_mode_Remote*, *leisure_screen_hours*, *gender_Male*, dan *occupation_Student*. *Sleep Hours* berada pada urutan kedelapan, kemudian diikuti *Work Screen Hours* dan *Social Hours per Week*, sebagaimana ditunjukkan pada Gambar 5. Nama kategori tersebut mengikuti keluaran *one-hot encoding* yang digunakan model.



Gambar 5 Sepuluh fitur terpenting XGBoost teroptimasi pada *seed* 42

Kedua model pada ilustrasi ini menempatkan durasi penggunaan layar dan kualitas tidur sebagai sumber informasi prediktif utama, tetapi berbeda pada urutan fitur setelahnya. Nilai *feature importance* menunjukkan kontribusi relatif fitur dalam masing-masing model, bukan arah hubungan atau bukti sebab-akibat terhadap stres. Urutan yang ditampilkan juga terbatas pada model *seed* 42 dan tidak merupakan rata-rata peringkat fitur dari 21 pembagian data.

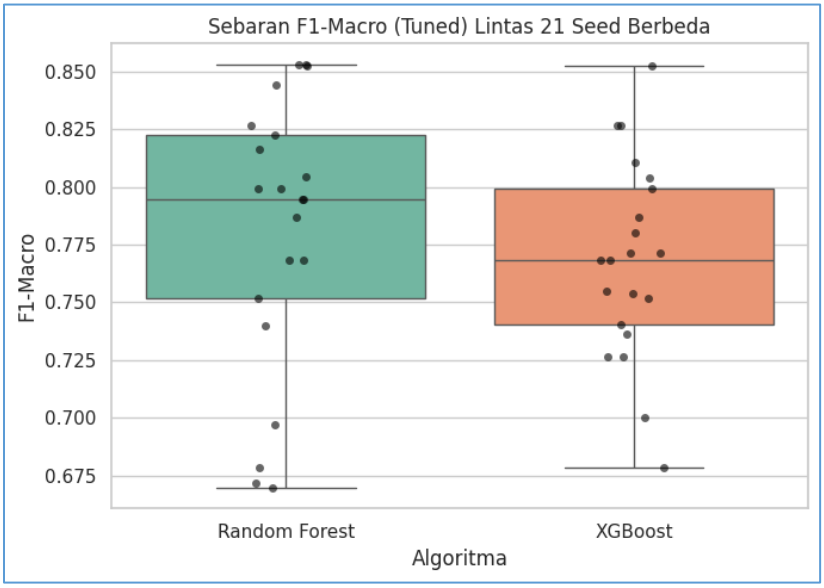
4.5 Uji Ketahanan Kinerja Model pada Berbagai Pembagian Data

Perbandingan berpasangan pada 21 *seed* menunjukkan bahwa *Random Forest* teroptimasi memperoleh *F1-Macro* lebih tinggi pada 11 pembagian, XGBoost pada sembilan pembagian, dan satu

 ISSN:2302-8149
e-ISSN:2540-9719

pembagian menghasilkan nilai yang sama. Rata-rata selisih *F1-Macro Random Forest* dikurangi XGBoost sebesar +0,0123. Dengan demikian, arah perbedaan kinerja tidak selalu sama pada setiap pembagian data.

Sebaran *F1-Macro* pada Gambar 6 memperlihatkan rentang kedua model yang saling tumpang tindih. Simpangan baku *Random Forest* sebesar 0,0595, sedangkan XGBoost sebesar 0,0431. XGBoost menunjukkan variasi *F1-Macro* yang lebih kecil pada kumpulan pembagian ini, meskipun rata-ratanya lebih rendah daripada *Random Forest*.



Gambar 6 Sebaran *F1-Macro* model teroptimasi pada 21 *seed*

Pada *seed* 42, *F1-Macro Random Forest* sebesar 0,8529 lebih tinggi daripada rata-ratanya sebesar 0,7807. XGBoost memperoleh 0,7714 pada pembagian tersebut, mendekati rata-ratanya sebesar 0,7684. Perbedaan antara hasil ilustratif dan hasil agregat ini menjadi alasan untuk menggunakan rata-rata serta variasi dari seluruh pembagian sebagai dasar utama interpretasi.

4.6 Uji Signifikansi Statistik Perbedaan Kinerja Model

Perbedaan *F1-Macro* model teroptimasi dianalisis menggunakan *Wilcoxon signed-rank test* pada pasangan hasil 21 *seed* serta *5×2cv paired t-test* yang dijalankan terpisah dengan pembagian 50:50. Ringkasan pengujian pada taraf signifikansi 0,05 disajikan pada Tabel 9.

Tabel 9 Hasil uji signifikansi kinerja *random forest* dan XGBoost

Metode Pengujian	Dasar Perbandingan	*p-value*	Hasil pada α = 0,05
Wilcoxon signed-rank test	21 pasangan hasil pembagian 80:20	0,2959	Tidak signifikan; indikatif
5×2cv paired t-test	10 pengukuran dari 5 replikasi dua arah dengan pembagian 50:50	0,5747	Tidak signifikan

Catatan: hasil Wilcoxon bersifat indikatif karena pembagian berulang menggunakan dataset yang sama sehingga observasi dapat muncul kembali dan hasil antarpembagian tidak sepenuhnya independen.

Wilcoxon menghasilkan p = 0,2959. Pada pengujian 5×2cv, selisih *F1-Macro* mengarah pada *Random Forest* dalam delapan dari sepuluh pengukuran, dengan rata-rata selisih +0,0290. Statistik uji sebesar t = −0,5999 dengan lima derajat kebebasan menghasilkan p = 0,5747. Kedua pengujian belum memberikan bukti yang cukup untuk menyatakan perbedaan kinerja yang signifikan pada taraf 0,05.

 ISSN:2302-8149
e-ISSN:2540-9719

Hasil ini membatasi klaim keunggulan *Random Forest* pada kecenderungan deskriptif dalam eksperimen yang dilakukan dan tidak membuktikan bahwa kedua algoritma setara.

4.7 Pembahasan

Dalam perbandingan kedua algoritma, *Random Forest* teroptimasi memperoleh rata-rata *F1-Macro* lebih tinggi daripada XGBoost, yaitu 0,7807 dibandingkan 0,7684. Namun, XGBoost menunjukkan variasi kinerja yang lebih kecil, dengan simpangan baku 0,0431 dibandingkan 0,0595 pada *Random Forest*. *Random Forest* memperoleh *F1-Macro* lebih tinggi pada 11 dari 21 pembagian data, sehingga keunggulannya tidak muncul pada seluruh pembagian. Hasil Wilcoxon dan *5×2cv paired t-test* pada Tabel 9 juga belum memberikan bukti perbedaan yang signifikan. Temuan ini menunjukkan kecenderungan keunggulan rata-rata *Random Forest* dalam eksperimen yang dilakukan, tanpa membuktikan bahwa salah satu algoritma selalu lebih baik atau bahwa keduanya setara.

Arah perbandingan tersebut sejalan dengan Fahria dkk. [10], yang melaporkan kinerja *Random Forest* lebih tinggi daripada XGBoost pada klasifikasi stres berbasis perilaku pengguna media sosial. Perbedaannya, penelitian tersebut menggunakan parameter bawaan, sedangkan penelitian ini membandingkan kondisi sebelum dan sesudah optimasi. Tamboto dan Jaya [11] juga melaporkan rata-rata hasil validasi silang *Random Forest* yang lebih tinggi pada klasifikasi stres berbasis parameter fisiologis tidur. Kesamaan ini terbatas pada arah temuan; perbedaan dataset, fitur, jumlah kelas, dan skema evaluasi membuat angka kinerja antarpenelitian tidak dapat dibandingkan secara langsung.

Pengaruh optimasi *hyperparameter* terlihat pada peningkatan rata-rata *F1-Macro* dan *recall* kelas "Aman" pada kedua algoritma, sebagaimana ditunjukkan pada Tabel 6. Namun, peningkatan tersebut tidak berlaku untuk seluruh metrik: rata-rata akurasi *Random Forest* sedikit menurun dari 0,8774 menjadi 0,8744. Perubahan ini sesuai dengan penggunaan *F1-Macro* sebagai dasar pemilihan konfigurasi GridSearchCV, yang memberikan bobot setara pada kedua kelas. Dengan demikian, hasil optimasi terutama menunjukkan perbaikan pengenalan kelas minoritas dalam konfigurasi yang diuji, tanpa menjamin peningkatan akurasi maupun kestabilan pada setiap pembagian data.

Dari sisi kontribusi fitur, Gambar 4 dan Gambar 5 menunjukkan bahwa Screen Time Hours dan Sleep Quality menempati dua urutan teratas pada kedua model teroptimasi untuk *seed* 42. Pada *Random Forest*, urutan berikutnya adalah Leisure Screen Hours dan Sleep Hours. Pada XGBoost, kategori mode kerja *In-person* dan *Remote* berada setelah dua fitur utama, kemudian diikuti Leisure Screen Hours, kategori jenis kelamin laki-laki, dan pekerjaan mahasiswa. Perbedaan tersebut menunjukkan bahwa kedua model memanfaatkan informasi tambahan dengan pembobotan yang berbeda. Urutan ini berlaku pada model yang ditampilkan dan tidak merupakan rata-rata peringkat fitur dari seluruh pembagian data.

Relevansi durasi penggunaan layar juga terlihat pada penelitian Fahria dkk. [10], yang menempatkan Daily Screen Time sebagai salah satu fitur penting. Hafin dkk. [14] melaporkan kontribusi durasi tidur dan penggunaan media sosial dalam prediksi risiko depresi pada remaja. Meskipun berbeda populasi dan target prediksi, temuan tersebut memberikan konteks mengenai penggunaan fitur perilaku digital dan tidur dalam pemodelan kesehatan mental. Nilai *feature importance* dalam penelitian ini tetap diinterpretasikan sebagai kontribusi relatif terhadap prediksi model, bukan sebagai bukti hubungan sebab-akibat atau mekanisme yang mendasari stres.

Perlakuan terhadap fitur proksi menjadi perbedaan lain dalam rancangan penelitian. Fahria dkk. [10] mempertahankan Happiness Index sebagai prediktor, sedangkan penelitian ini mengeluarkan Mental Wellness Index dan Productivity berdasarkan hasil diagnostik yang disajikan pada Subbab 4.1. Penghapusan tersebut membatasi pemodelan pada fitur demografis dan perilaku yang menjadi fokus penelitian. Korelasi tinggi mendukung kecurigaan terhadap kedekatan informasi kedua fitur dengan target, tetapi tidak dengan sendirinya membuktikan kebocoran data. Karena tidak dilakukan perbandingan khusus dengan dan tanpa kedua fitur, hasil penelitian tidak digunakan untuk menyimpulkan besarnya pengaruh penghapusan fitur terhadap kinerja.

SMOTENC digunakan untuk menangani data yang memuat variabel numerik dan kategorikal sebelum *one-hot encoding*, dengan pembentukan sampel sintetis terbatas pada data latih di setiap *fold*. Model *baseline* dan teroptimasi menggunakan prosedur penyeimbangan yang sama, sehingga perbandingan keduanya menelaah pengaruh optimasi dalam rancangan tersebut. Hasil ini tidak mengukur manfaat SMOTENC secara terpisah. Sebagai konteks, Sifaunnufus dkk. [13] memperoleh konfigurasi terbaik tanpa SMOTE pada klasifikasi stres berbasis EEG. Perbedaan tersebut

memperlihatkan bahwa manfaat penyeimbangan kelas perlu dipahami sesuai karakteristik data dan konfigurasi pemodelannya.

4.8 Keterbatasan Penelitian

Penelitian ini menggunakan satu dataset publik dari Kaggle yang mencakup 400 observasi, dengan 59 observasi pada kelas "Aman". Ukuran data dan keterbatasan informasi mengenai proses pengumpulan serta keterwakilan sampel membatasi generalisasi hasil. Ambang skor stres ≥ 6 merupakan definisi operasional klasifikasi dalam penelitian ini, bukan batas diagnosis klinis yang tervalidasi. Oleh karena itu, hasil prediksi perlu dipahami sesuai label yang dibentuk dari ambang tersebut.

Evaluasi pada 21 pembagian data menggunakan kembali observasi dari dataset yang sama, sehingga hasil antarpembagian tidak sepenuhnya independen dan uji Wilcoxon diperlakukan sebagai bukti indikatif. Pengujian berulang tersebut juga belum menggantikan validasi eksternal. Karena model belum diuji pada dataset independen, kinerjanya pada populasi atau lingkungan dengan karakteristik berbeda masih belum diketahui. Temuan penelitian ini karena itu terbatas pada perbandingan model dalam dataset dan rancangan evaluasi yang digunakan.

5 Kesimpulan

Random Forest teroptimasi menghasilkan rata-rata *F1-Macro* lebih tinggi daripada XGBoost dalam klasifikasi stres berbasis perilaku digital dan tidur pada dataset yang digunakan, tetapi kedua pengujian statistik tidak menunjukkan perbedaan yang signifikan. Optimasi *hyperparameter* meningkatkan rata-rata *F1-Macro Random Forest* dari 0,7721 menjadi 0,7807 dan XGBoost dari 0,7370 menjadi 0,7684, disertai peningkatan pengenalan kelas minoritas. Analisis *feature importance* pada model ilustratif menempatkan Screen Time Hours dan Sleep Quality sebagai dua fitur terpenting pada kedua algoritma, tanpa menunjukkan hubungan sebab-akibat terhadap stres. Temuan ini dibatasi oleh ukuran dan asal dataset, penggunaan ambang klasifikasi operasional, pembagian data berulang yang tidak sepenuhnya independen, serta belum adanya validasi eksternal. Penelitian lanjutan disarankan melakukan validasi pada dataset independen yang lebih besar dan beragam serta mengevaluasi ketahanan hasil terhadap pilihan ambang klasifikasi.

Referensi

[1] K. A. Devi and S. K. Singh, "*The Hazards of Excessive Screen Time: Impacts on Physical Health, Mental Health, and Overall well-being*," *J. Educ. Health Promot.*, Vol. 12, No. 1, Art. No. 413, Nov. 2023, DOI: 10.4103/jehp.jehp_447_23.

[2] R. M. S. Santos *et al.*, "*The Associations between Screen Time and Mental Health in Adults: A Systematic Review*," *J. Technol. Behav. SCI.*, Vol. 9, pp. 825–845, Feb. 2024, DOI: 10.1007/s41347-024-00398-7.

[3] A. Khan, G. McLeod, T. Hidajat, and E. J. Edwards, "*Excessive Smartphone use is Associated with Depression, Anxiety, Stress, and Sleep Quality of Australian Adults*," *J. Med. Syst.*, Vol. 47, No. 1, Art. No. 109, Oct. 2023, DOI: 10.1007/s10916-023-02005-3.

[4] M. Belal, N. Luong, T. Aledavood, and J. Kulshrestha, "*Examining the Association between Internet use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data with Questionnaires*," *J. Med. Internet Res.*, Vol. 28, Art. No. e78775, Jan. 2026, DOI: 10.2196/78775.

[5] C. Pieh *et al.*, "*Smartphone Screen Time Reduction Improves Mental Health: A Randomized Controlled Trial*," *BMC Med.*, Vol. 23, No. 1, Art. No. 107, Feb. 2025, DOI: 10.1186/s12916-025-03944-z.

[6] C. Zhong, M. Masters, S. M. Donzella, W. R. Diver, and A. V. Patel, "*Electronic Screen Use and Sleep Duration and Timing in Adults*," *JAMA Netw. Open*, Vol. 8, No. 3, Art. No. e252493, Mar. 2025, DOI: 10.1001/jamanetworkopen.2025.2493.

[7] T. O. Andersen *et al.*, "*Nighttime Smartphone Use, Sleep Quality, and Mental Health: Investigating a Complex Relationship*," *Sleep*, Vol. 46, No. 12, Art. No. zsad256, Sep. 2023, DOI: 10.1093/sleep/zsad256.

[8] M. C. Chandler, O. K. Ellison, A. L. McGowan, K. M. Fenn, and M. B. Pontifex, "*Physical Activity and Sleep Moderate the Relationship between Stress and Screen Time in College-Aged Adults*," *Journal of American College Health*, Vol. 72, No. 5, pp. 1401–1411, Jul. 2024, DOI: 10.1080/07448481.2022.2077110.

[9] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "*Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches*," *Sensors*, Vol. 24, No. 2, Art. No. 348, Jan. 2024, DOI: 10.3390/s24020348.

[10] D. N. Fahria, K. D. Tania, and R. D. Kurnia, "Analisis Komparatif Algoritma *Random Forest, Xgboost*, dan *Catboost* untuk Klasifikasi Tingkat Stres Pengguna Media Sosial," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, Vol. 11, No. 1, pp. 1843–1853, Jan. 2026, DOI: 10.36341/rabit.v11i1.7449.

[11] S. A. Tamboto and S. Jaya, "*Benchmarking XGBoost and Random Forest for Stress Level Prediction based on Sleep Physiological Parameters*," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, Vol. 6, No. 3, pp. 1501–1510, Jun. 2026, Accessed: Oct. 22, 2026. [Online]. Available: https://journal.irpi.or.id/index.php/malcom/article/view/2783

[12] S. W. Sugiharto and Y. Nataliani, "*Machine Learning Model Optimization and Interpretability Analysis for Classifying Student Stress Levels*," *Sistemasi: Jurnal Sistem Informasi*, Vol. 15, No. 5, pp. 1594–1607, May 2026, DOI: 10.32520/stmsi.v15i5.6239.

[13] F. I. S. Ms, F. A. Bachtiar, and E. S. Pramukantoro, "Perbandingan Klasifikasi Stres berbasis *Electroencephalography* dengan Kombinasi Ekstraksi dan Seleksi Fitur menggunakan Metode Tunggal dan *Ensemble*," *Jurnal Teknologi Informasi dan Ilmu Komputer*, Vol. 13, No. 3, pp. 713–722, Jun. 2026, Accessed: Oct. 22, 2026. [Online]. Available: https://jtiik.ub.ac.id/index.php/jtiik/id/article/view/10457/1604

[14] A. F. Hafin, N. T. Astuti, and S. Zahroh, "Analisis Prediksi Risiko Depresi pada Remaja menggunakan *Random Forest* dan *Logistic Regression* berdasarkan Pola Penggunaan Media Sosial," *JIIFKOM (Jurnal Ilmiah Informatika dan Komputer)*, Vol. 5, No. 2, pp. 67–84, Jul. 2026, DOI: 10.51901/jiifkom.v5i02.735.

[15] D. Pratama Putra, A. Apriyadi, Z. Firmansyah, K. Ditha Tania, and D. Kurniawan, "Perbandingan Model Regresi *Machine Learning* untuk Prediksi Skor Tingkat Stres berdasarkan Pola *Screen Time* Tahun 2025," *Jurnal Pendidikan dan Teknologi Indonesia*, Vol. 6, No. 4, pp. 733–741, Apr. 2026, DOI: 10.52436/1.jpti.1537.

[16] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "*Imbalanced Data Problem in Machine Learning: A Review*," *IEEE Access*, Vol. 13, pp. 13686–13699, Jan. 2025, DOI: 10.1109/ACCESS.2025.3531662.

[17] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "*Enhancing SMOTE for Imbalanced Data with Abnormal Minority Instances*," *Machine Learning with Applications*, Vol. 18, Art. No. 100597, Dec. 2024, DOI: 10.1016/j.mlwa.2024.100597.

[18] A. A. Huang and S. Y. Huang, "*Computation of the Distribution of Model Accuracy Statistics in Machine Learning: Comparison between Analytically Derived Distributions and Simulation-based Methods*," *Health SCI. Rep.*, Vol. 6, No. 4, Art. No. e1214, Apr. 2023, DOI: 10.1002/hsr2.1214.

[19] R. D. Riley and G. S. Collins, "*Stability of Clinical Prediction Models Developed using Statistical or Machine Learning Methods*," *Biometrical Journal*, Vol. 65, No. 8, Art. No. 2200302, Dec. 2023, DOI: 10.1002/bimj.202200302.

[20] A. Apicella, F. Isgrò, and R. Prevete, "*Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning*," *Artif. Intell. Rev.*, Vol. 58, No. 11, Art. no. 339, Nov. 2025, DOI: 10.1007/s10462-025-11326-3.

[21] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "*Data Leakage Inflates Prediction Performance in Connectome-based Machine Learning Models*," *Nature Communications* , Vol. 15, No. 1, Art. No. 1829, Feb. 2024, DOI: 10.1038/s41467-024-46150-w.

 ISSN:2302-8149
e-ISSN:2540-9719

[22] S. Kapoor and A. Narayanan, "*Leakage and the Reproducibility Crisis in Machine-Learning-based Science*," *Patterns*, Vol. 4, No. 9, Art. No. 100804, Sep. 2023, DOI: 10.1016/j.patter.2023.100804.
[23] L. Breiman, "*Random Forests*," *Mach. Learn.*, Vol. 45, No. 1, pp. 5–32, Oct. 2001, DOI: 10.1023/A:1010933404324.
[24] T. Chen and C. Guestrin, "*XGBoost: A Scalable Tree Boosting System*," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.
[25] T. G. Dietterich, "*Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*," *Neural Comput.*, Vol. 10, No. 7, pp. 1895–1923, Oct. 1998, DOI: 10.1162/089976698300017197.