Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

Evaluasi Prediksi Stres Berdasarkan Tidur dan Kognitif Menggunakan Machine Learning

Evaluating Stress Prediction from Sleep and Cognitive Indicators Using Machine Learning

Ibnu Sina Maulana*, Hafiyyan Putra Pratama
Kampus UPI di Purwakarta, Program Studi Sistem Telekomunikasi, Universitas Pendidikan Indonesia, Jawa Barat, Indonesia
Email: 1,*ibnusinam@upi.edu, 2hafiyyan@upi.edu
Email Penulis Korespondensi: ibnusinam@upi.edu
Received: **01/06/2026**, Revised: **20/10/2026**, Accepted: **28/10/2026**, Published: **31/10/2026**

Abstrak−Indikator tidur dan kognitif memiliki relevansi teoretis terhadap stres, tetapi relevansi tersebut belum menjamin prediksi individual pada data kecil. Studi eksploratif ini mengevaluasi keterpelajaran skor *Stress_Level* pada dataset publik *Sleep Deprivation and Cognitive Performance* yang memuat 60 observasi dan 12 prediktor. Statistik deskriptif dan korelasi dilengkapi validasi silang bertingkat dengan lima *fold* luar, tiga pengulangan, serta tiga *fold* dalam. Pengodean, standardisasi, dan seleksi fitur dipelajari hanya pada data latih. Enam keluarga regresi dibandingkan dengan *baseline* rata-rata, *least absolute shrinkage and selection operator* (*LASSO*), dan *recursive feature elimination* (*RFE*) berbasis *Ridge*. Seluruh interval kepercayaan 95% korelasi marginal melintasi nol; korelasi terbesar adalah kualitas tidur dengan koefisien 0,190. Seluruh model berbasis fitur menghasilkan rata-rata koefisien determinasi negatif, dari −0,470 hingga −0,235, yang menunjukkan kegagalan generalisasi pada evaluasi internal ini. *Baseline* mencapai galat absolut rata-rata terendah sebesar 9,65 ± 1,33, dibandingkan dengan *RFE–Ridge* 9,69 ± 1,35 dan *Ridge* 9,76 ± 1,41; angka setelah tanda ± adalah standar deviasi antar-*fold* luar. Rata-rata prediksi *Ridge* dari pengujian silang terkonsentrasi pada 14,67–19,72, sedangkan target mencakup 0–40. Tuning dan seleksi fitur belum menunjukkan nilai tambah prediktif dibandingkan *baseline*. Temuan negatif ini memberikan diagnosis empiris tentang keterbatasan informasi dataset dan tidak mendukung penerapan model untuk skrining atau diagnosis stres. Ukuran sampel kecil, dokumentasi instrumen terbatas, dan ketiadaan validasi eksternal membatasi kesimpulan. Pengembangan berikutnya memerlukan pengukuran stres tervalidasi, data longitudinal, dan fitur fisiologis serta konteks psikososial yang diselaraskan secara temporal.

Kata Kunci: Stres; Kognitif; Regresi; Validasi Silang Bertingkat; Studi Eksploratif

Abstract−Sleep and cognitive indicators are theoretically relevant to stress, but this relevance does not guarantee individual prediction in small datasets. This exploratory study evaluates the learnability of Stress_Level in the public Sleep Deprivation and Cognitive Performance dataset containing 60 observations and 12 predictors. Descriptive analysis is complemented by nested cross-validation with five outer folds, three repetitions, and three inner folds. Encoding, scaling, and feature selection are fitted exclusively on training data. Six regression families are compared with a training-mean baseline, least absolute shrinkage and selection operator (LASSO), and recursive feature elimination (RFE) with Ridge. All marginal 95% correlation intervals include zero; the largest correlation concerns sleep quality (0.190). All feature-based models yield negative mean coefficients of determination, ranging from −0.470 to −0.235, indicating failure to generalize in this internal evaluation. The baseline achieves the lowest mean absolute error of 9.65 ± 1.33, compared with 9.69 ± 1.35 for RFE–Ridge and 9.76 ± 1.41 for Ridge; values after ± represent standard deviations across outer folds. Averaged cross-validated Ridge predictions concentrate within 14.67–19.72 despite the target range of 0–40. Tuning and feature selection do not demonstrate added predictive value over the baseline. These negative findings diagnose the information limits of this dataset and do not support stress screening or diagnosis. Small sample size, limited instrument documentation, and absent external validation restrict inference. Future studies require validated stress measurements, longitudinal observations, and temporally aligned physiological and psychosocial features.

Keywords: Stress; Cognition; Regression; Nested Cross-Validation; Exploratory Study

1. PENDAHULUAN

Tidur mendukung kewaspadaan, memori, dan pengendalian respons. Tinjauan terhadap dewasa muda menunjukkan bahwa dampak pembatasan tidur bergantung pada bentuk deprivasi dan luaran neurobehavioral yang diukur [1]. Bukti sintesis juga menghubungkan pembatasan tidur dengan pembentukan memori [2], kontrol inhibisi [3], dan perubahan pengalaman emosional [4]. Tinjauan National Sleep Foundation tahun 2026 mempertahankan konteks rekomendasi durasi tidur dewasa sekitar 7–9 jam dan menekankan perbedaan modalitas pengukurannya [5]. Pada dataset yang dianalisis, rata-rata durasi tidur hanya 5,81 jam, tetapi korelasi durasi tidur dengan skor stres mendekati nol. Perbedaan antara relevansi teoretis dan informasi yang teramati ini memotivasi evaluasi prediksi yang memeriksa apakah fitur benar-benar membantu di luar prediksi konstan.

Stres mencakup penilaian subjektif terhadap tuntutan, sumber daya *coping*, dan konteks kehidupan. Huang dkk. [6] memeriksa hubungan *perceived stress* dan kualitas tidur melalui model mediasi serta moderasi, sehingga hubungan tersebut tidak cukup direduksi menjadi satu ukuran durasi. Kafein dipilih karena berkaitan dengan karakteristik tidur [7], sedangkan aktivitas fisik memberikan konteks perilaku yang juga berhubungan dengan tidur [8]. Usia, gender, dan *BMI* memberi konteks demografis serta antropometris. Skor regulasi emosi dipisahkan secara konseptual dari *Stroop*, *N-Back*, dan *PVT* karena domain penilaiannya berbeda. Variabel tersebut memiliki dasar untuk diuji, tetapi tidak diasumsikan sebagai penyebab stres atau jaminan prediksi yang akurat.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 51
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

Tutunji dkk. [9] menggabungkan biosensor *wearable* dan *ecological momentary assessment* (*EMA*) pada 83 mahasiswa selama minggu ujian dan minggu kontrol. Analisis model campuran serta *machine learning* memperlihatkan bahwa gabungan *EMA* dan fisiologi menurunkan galat klasifikasi dibandingkan masukan tunggal, sementara model individual lebih baik daripada model kelompok. Keterbatasan relevansinya bagi penelitian ini adalah perbedaan target: klasifikasi konteks minggu ujian tidak setara dengan regresi skor stres. Temuan tersebut tetap menegaskan bahwa sinyal fisiologis perlu ditafsirkan bersama konteks psikologis dan perbedaan individu.

Kinoshita dkk. [10] mengikuti 40 pekerja selama delapan minggu dengan *wearable*, *EMA*, dan penilaian stres serta kesejahteraan. Analisis multilevel menemukan hubungan variasi waktu tidur dengan stres dan hubungan durasi tidur sebelumnya dengan kantuk berikutnya. Rancangan longitudinalnya membedakan variasi intraindividu dari perbedaan antarindividu, tetapi ukuran sampel pilot dan populasi pekerja membatasi generalisasi. Berbeda dari rancangan itu, dataset penelitian ini tidak menyediakan urutan hari atau pengukuran berulang; ketiadaan dokumentasi waktu tersebut membatasi interpretasi mekanisme yang mungkin melandasi prediksi.

Yao dkk. [11] memakai aplikasi *EMA* pada 33 partisipan dengan pencatatan lokasi untuk menelaah konteks stresor harian. Analisis tematik menunjukkan bahwa beberapa jenis stresor berbeda menurut lokasi, sedangkan sebagian lain muncul lintaslingkungan. Fokusnya adalah pemetaan konteks, bukan estimasi performa regresi universal; ukuran sampelnya juga membatasi keterwakilan populasi. Studi ini menunjukkan jenis informasi psikososial yang tidak tersedia dalam tabel tidur dan kognitif, sehingga penambahan algoritma saja belum tentu mengatasi keterbatasan representasi fitur.

Vos dkk. [12] meninjau generalisasi *machine learning* untuk pemantauan stres berbasis *wearable*, dengan perhatian pada protokol label, validasi, dan reproduksibilitas. Temuan lintasstudi menunjukkan bahwa generalisasi pada data baru tetap menjadi kendala; heterogenitas dataset dan evaluasi membatasi perbandingan langsung antarangka performa. Razavi dkk. [13] menelaah 98 publikasi dan memetakan penggunaan sinyal fisiologis, *preprocessing*, serta keluarga algoritma. Tinjauan tersebut menemukan banyak penggunaan ukuran denyut jantung dan respons kulit, tetapi juga mengidentifikasi kebutuhan interpretabilitas, personalisasi, dan pengujian naturalistik. Kedua tinjauan memberi dasar memilih keluarga model yang beragam, namun tidak membuktikan bahwa keluarga tertentu akan berhasil pada dataset tabular kecil ini.

Kesenjangan teknis penelitian ini adalah penilaian nilai tambah indikator yang tersedia ketika evaluasi dibatasi oleh sampel kecil dan dokumentasi pengukuran yang terbatas. Literatur kebocoran data menunjukkan bahwa penggunaan informasi evaluasi dalam *preprocessing* atau seleksi fitur dapat menghasilkan angka terlalu optimistis [14], [15], [16]. Karena itu, pembanding konstan, transformasi di dalam *pipeline*, dan pemisahan *tuning* dari evaluasi luar menjadi bagian inti desain [17]. Penelitian ini tidak mengklaim sebagai penemu pertama protokol tersebut; kontribusinya berupa penerapan yang transparan untuk memeriksa keterpelajaran target dan mendiagnosis penyusutan rentang prediksi pada dataset tertentu.

Dengan 60 observasi dan 12 prediktor, rasio observasi terhadap prediktor hanya lima, sebelum kompleksitas model diperhitungkan. Sampel tersebut layak untuk audit dan pemeriksaan kelayakan secara eksploratif, tetapi tidak dijadikan pembenaran daya statistik atau kesiapan klinis. Acuan TRIPOD+AI digunakan untuk memperjelas sumber data, pelatihan, evaluasi, dan batas interpretasi [18]. Tujuan penelitian adalah mengaudit dataset, mengestimasi hubungan marginal fitur dengan *Stress_Level*, membandingkan regresi terhadap *baseline* melalui validasi silang bertingkat, serta menguji pengaruh regularisasi dan seleksi fitur. Analisis residu melengkapi metrik untuk menunjukkan pola kesalahan. Kontribusi penelitian berupa bukti empiris batas keterpelajaran dan prosedur evaluasi yang dapat diperiksa, tanpa klaim biomarker, sebab-akibat, atau alat diagnosis.

2. METODOLOGI PENELITIAN

2.1 Desain dan Tahapan Penelitian

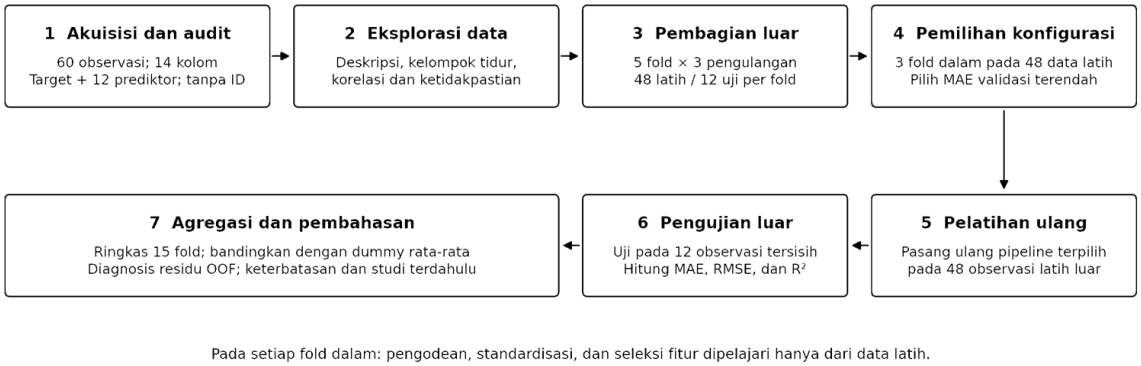
Penelitian ini menggunakan desain kuantitatif eksploratif melalui analisis sekunder dataset publik. Metode utamanya adalah komparasi regresi terawasi dengan validasi silang bertingkat untuk menilai apakah 12 prediktor memberikan galat lebih rendah daripada *baseline* rata-rata data latih. Unit analisis adalah satu rekaman observasi, sedangkan *Stress_Level* diperlakukan sebagai target numerik. Enam keluarga regresi, *LASSO*, dan *RFE–Ridge* diuji pada partisi yang sama. Gambar 1 memperlihatkan urutan penelitian dan pemisahan proses pemilihan konfigurasi dari pengujian model.

Tahap pertama adalah memperoleh dan mengaudit dataset, menetapkan target serta prediktor, dan memeriksa nilai hilang, duplikasi, serta dokumentasi variabel. Tahap kedua adalah analisis deskriptif, pengelompokan durasi tidur, dan estimasi korelasi. Hasil eksplorasi digunakan untuk memahami karakteristik data, tanpa menghapus prediktor berdasarkan korelasi. Tahap ketiga membagi data melalui lima *fold* luar yang diulang tiga kali; pada setiap partisi, 48 observasi menjadi data latih dan 12 observasi disisihkan sebagai data uji luar.

Tahap keempat memilih konfigurasi *pipeline* pengodean, standardisasi, seleksi fitur bila digunakan, dan estimator melalui tiga *fold* dalam pada 48 observasi latih, berdasarkan MAE validasi terendah. Tahap kelima melatih ulang *pipeline* terpilih pada seluruh data latih luar. Tahap keenam menguji *pipeline* dan *baseline* pada 12 observasi uji luar; prosedur diulang hingga diperoleh 15 hasil evaluasi. Tahap ketujuh merangkum MAE, RMSE, dan R^2, menelaah pemilihan fitur, serta mendiagnosis residu dan penyusutan rentang prediksi. Analisis sensitivitas pembagian 80:20 dilaporkan terpisah dari evaluasi utama.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 52
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 1. Tahapan penelitian dengan komparasi regresi terawasi dan validasi silang bertingkat

2.2 Sumber Data dan Batas Dokumentasi

Data berasal dari *Sleep Deprivation and Cognitive Performance* pada Kaggle (doi: 10.34740/KAGGLE/DSV/10569288) [19]. Evaluasi ulang memakai berkas ARFF salinan OpenML ID 46754 versi 1 (https://www.openml.org/d/46754), berisi 60 baris dan 14 kolom. Konsistensi sumber diperiksa melalui nama kolom, statistik deskriptif, jumlah kelompok tidur, dan hasil pemodelan acuan.

Penggunaan domain indikator tidur mempertimbangkan penelitian Larasati dkk. [20] tentang klasifikasi stres berdasarkan kondisi tidur. Perbedaan target dan dataset dipertimbangkan secara eksplisit pada Subbagian 3.9.

Dokumentasi repositori belum memungkinkan verifikasi rekrutmen, persetujuan etik studi primer, reliabilitas instrumen, periode rujukan skor, atau pembangkitan setiap kolom. Observasi diperlakukan sebagai rekaman dataset publik dengan keterwakilan populasi yang belum diketahui. Rentang *Stress_Level* 0–40 tidak cukup untuk mengidentifikasinya sebagai instrumen psikometri tertentu. Analisis sekunder ini tidak melibatkan perekrutan atau pengukuran baru.

Participant_ID dihapus dari prediktor karena merupakan pengenal. *Stress_Level* menjadi target numerik dan 12 kolom lainnya menjadi prediktor. Tidak dilakukan diskretisasi target menjadi kelas stres karena ambang klinisnya tidak terdokumentasi. Perlakuan skor sebagai target regresi merupakan asumsi operasional yang menyetarakan jarak antarnilai numerik; kesahihan psikometrik asumsi ini belum dapat diuji. Definisi kelompok dan satuan variabel dirangkum dalam Tabel 1.

Tabel 1. Variabel dataset dan peran analisis

Variabel	Makna dan batas dokumentasi	Peran
Participant_ID	Identitas observasi	Dikeluarkan
Sleep_Hours	Durasi tidur dalam jam	Tidur
Sleep_Quality_Score	Skor kualitas tidur; arah skor belum terverifikasi	Tidur
Daytime_Sleepiness	Skor kantuk siang; instrumen belum terverifikasi	Tidur
Stroop_Task_Reaction_Time	Waktu reaksi *Stroop*; satuan belum terverifikasi	Kognitif
N_Back_Accuracy	Akurasi *N-Back*; rentang numerik 50,90–99,73	Kognitif
PVT_Reaction_Time	Waktu reaksi *PVT*; satuan belum terverifikasi	Kognitif
Emotion_Regulation_Score	Skor regulasi emosi; arah dan instrumen belum terverifikasi	Psikologis
Age dan Gender	Usia numerik dan kategori gender dalam berkas	Demografis
BMI	Indeks massa tubuh dalam berkas	Antropometris
Caffeine_Intake	Nilai konsumsi kafein; dosis dan waktu tidak tersedia	Kebiasaan
Physical_Activity_Level	Skor aktivitas; durasi dan intensitas tidak tersedia	Kebiasaan
Stress_Level	Skor numerik target; rentang teramati 0–40	Target

2.3 Audit Data dan Analisis Eksploratif

Audit mencakup jumlah observasi, keunikan identitas, nilai hilang, dan duplikasi setelah identitas dikeluarkan. Tidak ditemukan nilai hilang atau duplikasi, sehingga tidak dilakukan imputasi maupun penghapusan observasi. Nilai ekstrem dipertahankan karena tidak ada bukti kesalahan pencatatan. Rata-rata, standar deviasi sampel, median, minimum, dan maksimum digunakan untuk mendokumentasikan karakteristik data.

Pengelompokan tidur menggunakan batas operasional <7 jam dan 7–9 jam, dengan mengacu pada rekomendasi durasi tidur dewasa [5]. Pengelompokan hanya digunakan untuk deskripsi, tidak dimasukkan sebagai fitur tambahan dan tidak dipakai untuk menyatakan gangguan tidur. Korelasi Pearson dihitung untuk seluruh prediktor numerik; gender dikodekan laki-laki = 1 dan perempuan = 0 untuk korelasi point-biserial yang ekuivalen. Interval kepercayaan 95% marginal dihitung menggunakan transformasi Fisher z dengan galat baku $1/\sqrt{(n-3)}$. Interval tersebut belum dikoreksi untuk perbandingan multipel dan tidak dimaksudkan sebagai uji konfirmatori.

Eksplorasi memakai seluruh data hanya untuk deskripsi, tanpa menyaring prediktor berdasarkan korelasi. Semua prediktor dipertahankan pada model tanpa seleksi; subset *LASSO* dan *RFE* dipelajari melalui loop dalam. Tahap eksplorasi pada Gambar 1 tidak menentukan subset bagi data uji. Evaluasi tetap bersifat *post hoc* dan internal.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 53
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

2.4 Prapemrosesan dan Validasi Silang Bertingkat

Evaluasi utama menggunakan *RepeatedKFold* lima *fold* luar, tiga pengulangan, dan *random_state* = 42. Setiap *fold* luar memuat 48 observasi latih dan 12 uji. *GridSearchCV* pada data latih luar memakai *KFold* tiga *fold* dengan *shuffle* = *True* dan seed 100 + k, untuk k = 0,…,14. Partisi dalam memuat 32 observasi latih dan 16 validasi. Konfigurasi dengan MAE validasi dalam terendah dipasang ulang pada 48 observasi latih lalu diuji pada 12 observasi uji luar. Semua model memakai partisi luar yang sama.

Pipeline menggabungkan *ColumnTransformer*, seleksi fitur bila digunakan, dan estimator. *StandardScaler* mempelajari rata-rata serta simpangan baku fitur numerik hanya pada *fold* latih; skala yang sebanding diperlukan terutama untuk model berpenalti [21]. Gender ditransformasikan dengan *OneHotEncoder*, *drop* = *if_binary*, dan *handle_unknown* = *ignore*. Dua kategori gender menghasilkan satu fitur biner sehingga total hasil transformasi tetap 12 fitur. Data validasi dan uji hanya menerima transformasi dari parameter data latih. *LASSO* atau *RFE* dipasang ulang di setiap pelatihan dalam, tanpa memakai label uji luar.

Risiko kebocoran informasi evaluasi ke pelatihan dibedakan dari variasi estimasi akibat sampel kecil [14], [15]. Validasi bertingkat memisahkan *tuning* dari pengujian luar, tanpa menambah informasi atau menjamin keterwakilan data [16], [17]. Pengodean dan standardisasi dipelajari hanya dari data latih melalui *pipeline*, sejalan dengan pemisahan prapemrosesan pada Ichwani dkk. [22].

Analisis sensitivitas memakai *train_test_split* 80:20, *shuffle* = *True*, dan *random_state* = 42, dengan 48 observasi latih dan 12 uji. Konfigurasi dasar pada pembagian ini dilaporkan terpisah dan tidak digunakan untuk memilih model akhir. Seluruh observasi berasal dari dataset yang sama, sehingga analisis sensitivitas bukan evaluasi eksternal.

2.5 Metode Regresi dan Seleksi Fitur

Linear Regression memodelkan hubungan aditif linier, sedangkan *Ridge* membatasi koefisien melalui regularisasi [21]. *Elastic Net* dan *LASSO* menguji regularisasi serta sparsitas. *SVR* RBF mewakili hubungan nonlinier halus, *Random Forest* menguji interaksi melalui *bagging*, dan *Gradient Boosting* membangun pohon bertahap. Perbandingan keluarga tersebut memeriksa apakah kegagalan prediksi terbatas pada asumsi linier.

RFE–Ridge menerapkan *recursive feature elimination* [23] memakai *Ridge* α = 1 sebagai pemeringkat, lalu *Ridge* akhir pada subset terpilih. Jumlah fitur dan α akhir dipilih di loop dalam. Koefisien *LASSO* dengan nilai absolut $\leq 10^{-8}$ dianggap nol hanya untuk pelaporan. *DummyRegressor* memakai rata-rata target data latih tanpa prediktor. Tabel 2 merinci ruang pencarian yang dievaluasi; ruang terbatas ini tidak diklaim menghasilkan optimum global.

Tabel 2. Konfigurasi dan ruang *tuning* pada loop dalam

Pendekatan	Konfigurasi yang diuji	Jumlah
Dummy rata-rata	strategy = mean	1
Linear Regression	*fit_intercept* = *True*	1
Ridge	α ∈ {0,1; 1; 10; 100; 1000}	5
Elastic Net	α ∈ {0,1; 1; 10}; *l1_ratio* ∈ {0,1; 0,5; 0,9}	9
SVR RBF	C ∈ {0,1; 1; 10}; ε ∈ {0,1; 1}; γ ∈ {scale; 0,01}	12
Random Forest	100 pohon; *max_depth* ∈ {2; *None*}; *min_samples_leaf* ∈ {2; 5}; *max_features* ∈ {0,5; 1,0}	8
Gradient Boosting	100 pohon; *learning_rate* ∈ {0,01; 0,05}; *max_depth* ∈ {1; 2}; *min_samples_leaf* ∈ {3; 6}	8
LASSO	α ∈ {0,1; 1; 5; 10}; *max_iter* = 20 000	4
RFE–Ridge	Fitur ∈ {4; 8; 12}; α akhir ∈ {1; 100; 1000}; α pemeringkat = 1	9

Random Forest dan *Gradient Boosting* memakai *random_state* = 42; *Elastic Net* serta *LASSO* menggunakan *max_iter* = 20 000. Parameter yang tidak tercantum pada Tabel 2 memakai nilai bawaan *scikit-learn* 1.8.0. Standardisasi numerik diterapkan seragam untuk konsistensi *pipeline*, meskipun pohon tidak memerlukannya. Implementasi menggunakan Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, dan *scikit-learn* 1.8.0. Ruang pencarian dibatasi agar dapat direproduksi dan tidak menjadi pencarian adaptif berulang terhadap hasil *fold* luar.

2.6 Metrik Evaluasi dan Diagnosis Residu

Kinerja model dinilai dengan galat absolut rata-rata (*mean absolute error*, MAE) pada Persamaan (1), akar galat kuadrat rata-rata (*root mean squared error*, RMSE) pada Persamaan (2), dan koefisien determinasi (R^2) pada Persamaan (3). MAE dan RMSE dilaporkan dalam satuan skor *Stress_Level*.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_{\mathrm{uji}}\right)^2} \tag{3}$$

dengan n jumlah observasi uji, y_i skor aktual, $\hat{y}_i$ prediksi, dan $\bar{y}_{\mathrm{uji}}$ rata-rata target *fold* uji. $R^2 < 0$ berarti kesalahan kuadrat melebihi prediksi rata-rata *fold* uji. Penyebut ini berbeda dari *baseline* operasional yang memakai rata-rata data latih; karena itu, R^2 dan perbandingan langsung terhadap dummy dilaporkan terpisah.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 54
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

Metrik diringkas sebagai rata-rata dan standar deviasi sampel atas 15 *fold* luar. Standar deviasi menggambarkan variasi antarpartisi, bukan standard error atau interval kepercayaan. Karena observasi berulang, *fold* tidak dianggap independen dan tidak diuji dengan uji t biasa. Pemeringkatan MAE digunakan secara deskriptif, tanpa klaim pemenang yang tervalidasi independen.

Residu didefinisikan $e_i = y_i - \hat{y}_i$. Setiap observasi memiliki tiga prediksi *out-of-fold* (OOF), masing-masing dari model yang tidak melatih observasi itu. Ketiga prediksi dirata-ratakan hanya untuk diagnosis sehingga tersedia 60 titik; agregasi ini tidak menggantikan metrik 15 *fold* luar. *Ridge* menjadi ilustrasi regularisasi, tanpa dipilih karena peringkat terbaik. Plot aktual–prediksi, residu–prediksi, histogram, dan Q–Q memeriksa rentang prediksi serta distribusi kesalahan.

Uji Shapiro–Wilk pada residu dan korelasi Spearman antara |residu| dan prediksi dilaporkan sebagai diagnosis eksploratif. Prediksi silang saling bergantung melalui data latih yang tumpang tindih, sehingga nilai p tidak diperlakukan sebagai inferensi konfirmatori. Untuk menggambarkan ketidakpastian perbandingan *Ridge* dengan dummy, dilakukan *bootstrap* berpasangan 5 000 kali pada 60 selisih kesalahan absolut dari prediksi OOF rata-rata. Interval persentil bersifat kondisional pada prediksi tersebut, tidak melatih ulang model, dan tidak mencakup seluruh ketidakpastian pemilihan maupun pelatihan *pipeline*.

3. HASIL DAN PEMBAHASAN

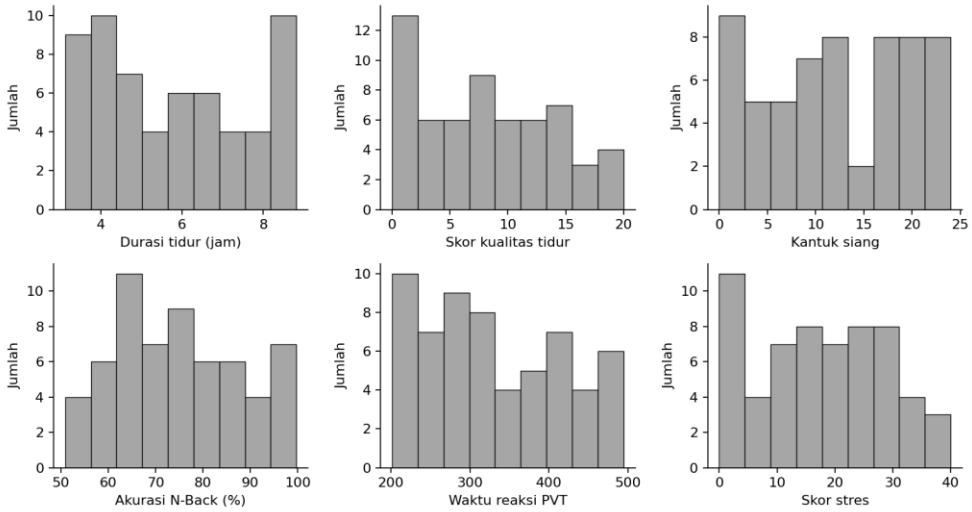
3.1 Hasil Audit dan Distribusi Data

Audit memastikan 60 identitas unik, 14 kolom, nol nilai hilang, dan nol duplikasi setelah *Participant_ID* dikeluarkan. Tabel 3 menyajikan rata-rata durasi tidur 5,81 jam dengan standar deviasi 1,83, serta target rata-rata 17,87 dengan standar deviasi 11,02. Kelengkapan numerik tidak membuktikan validitas pengukuran: seluruh kolom dapat terisi meskipun konstruk, satuan, atau reliabilitasnya belum terdokumentasi.

Tabel 3. Statistik deskriptif variabel numerik pada 60 observasi

Variabel	Rata-rata	SD	Minimum	Median	Maksimum
Durasi tidur	5,81	1,83	3,12	5,69	8,82
Kualitas tidur	8,32	5,64	0,00	8,00	20,00
Kantuk siang	12,00	7,58	0,00	11,50	24,00
Waktu reaksi *Stroop*	3,24	0,83	1,60	3,26	4,49
Akurasi *N-Back*	75,01	13,67	50,90	74,27	99,73
Regulasi emosi	38,15	17,13	10,00	37,00	67,00
Waktu reaksi *PVT*	332,54	87,80	201,56	327,21	494,55
Usia	29,52	8,17	18,00	28,50	43,00
BMI	27,33	4,54	18,74	27,37	34,93
Kafein	2,38	1,69	0,00	2,50	5,00
Aktivitas fisik	4,12	3,11	0,00	4,00	10,00
Skor stres	17,87	11,02	0,00	17,50	40,00

Gambar 2 menunjukkan rentang skor dan heterogenitas observasi. Distribusi target yang lebar memperbesar variasi nilai yang harus dicakup model, tetapi tidak secara otomatis menyebabkan R^2 negatif; kegagalan prediksi muncul ketika hubungan yang dipelajari tidak menekan kesalahan pada observasi yang tidak digunakan untuk pelatihan. Data ini juga tidak menyediakan durasi tindak lanjut atau pengukuran berulang. Oleh sebab itu, rendahnya tidur rata-rata hanya mendeskripsikan dataset dan tidak dapat dipakai untuk memperkirakan prevalensi kurang tidur pada populasi umum.



Gambar 2. Distribusi enam variabel utama

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 55
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

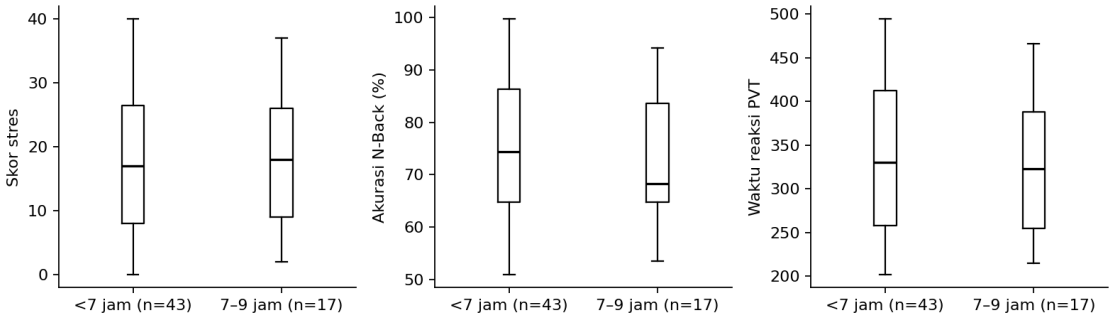
3.2 Perbandingan Deskriptif Kelompok Tidur

Sebanyak 43 observasi (71,67%) memiliki durasi tidur <7 jam dan 17 (28,33%) berada pada 7–9 jam. Tidak terdapat durasi >9 jam. Pada Tabel 4, skor stres kelompok tidur pendek justru sedikit lebih rendah, yaitu 17,65 dibanding 18,41, dengan selisih −0,76 poin. Waktu reaksi *PVT* berbeda 19,92 satuan numerik ke arah lebih lambat pada kelompok tidur pendek. Tanpa verifikasi satuan dan ketidakpastian perbedaan, angka tersebut tidak dapat diterjemahkan sebagai efek klinis.

Tabel 4. Rata-rata indikator berdasarkan kelompok durasi tidur

Kelompok	n	Stres	Kantuk	Kualitas tidur	*N-Back*	*PVT*
<7 jam	43	17,65	11,28	8,19	75,82	338,18
7–9 jam	17	18,41	13,82	8,65	72,95	318,26

Gambar 3 melengkapi rata-rata pada Tabel 4 dengan informasi sebaran. Arah perbedaan kantuk, *N-Back*, dan skor stres tidak seragam terhadap dugaan sederhana bahwa tidur lebih singkat selalu diikuti hasil yang lebih buruk. Kondisi ini membatasi narasi biologis yang dapat dibangun dari data. Perbedaan tersebut mungkin dipengaruhi variasi individu, konteks pengukuran, kesalahan ukur, atau struktur dataset; mekanismenya tidak dapat dipisahkan melalui perbandingan deskriptif ini. Pengelompokan juga menghilangkan variasi durasi kontinu sehingga tidak dipakai sebagai pembenaran hubungan kausal.



Gambar 3. Sebaran skor stres, akurasi *N-Back*, dan waktu reaksi *PVT* pada dua kelompok durasi tidur

3.3 Hubungan Marginal dan Ketidakpastiannya

Korelasi terbesar secara absolut adalah *Sleep_Quality_Score* dengan r = 0,190, disusul *BMI* sebesar 0,175. Untuk durasi tidur, korelasi hanya −0,021. Tabel 5 melaporkan seluruh 12 prediktor agar pemilihan beberapa fitur dengan korelasi terbesar tidak menimbulkan kesan adanya hubungan yang sudah terkonfirmasi. Semua interval kepercayaan marginal melintasi nol, termasuk kualitas tidur dengan interval −0,067 hingga 0,423. Rentang tersebut mencakup hubungan negatif lemah, nol, dan positif sedang; dataset belum mengestimasi arah dan besar hubungan secara presisi.

Tabel 5. Korelasi Pearson dengan *Stress_Level* dan interval Fisher z

Prediktor	r	IK 95% marginal	p eksploratif
Kualitas tidur	0,190	[−0,067; 0,423]	0,146
BMI	0,175	[−0,082; 0,411]	0,180
Kantuk siang	−0,165	[−0,403; 0,092]	0,206
Regulasi emosi	−0,148	[−0,387; 0,110]	0,259
Akurasi *N-Back*	−0,123	[−0,366; 0,135]	0,348
Waktu reaksi *PVT*	−0,118	[−0,361; 0,140]	0,370
Gender laki-laki	−0,079	[−0,326; 0,179]	0,550
Waktu reaksi *Stroop*	−0,049	[−0,299; 0,208]	0,713
Usia	0,046	[−0,211; 0,296]	0,729
Aktivitas fisik	0,041	[−0,215; 0,292]	0,755
Kafein	−0,021	[−0,273; 0,234]	0,874
Durasi tidur	−0,021	[−0,273; 0,235]	0,876

Gambar 4 memperlihatkan ketidakpastian yang tidak tampak ketika hanya nilai korelasi tunggal dilaporkan. Lemahnya korelasi marginal tidak membuktikan bahwa tidak ada hubungan nonlinier atau interaksi. Sebaliknya, hasil itu juga tidak memberi dasar untuk menyebut fitur sebagai sinyal awal yang berguna secara klinis. Tanda positif pada kualitas tidur tidak boleh diartikan bahwa tidur yang lebih baik meningkatkan stres karena arah skornya belum diverifikasi. Interpretasi psikologis harus menunggu dokumentasi instrumen dan replikasi pada data independen.

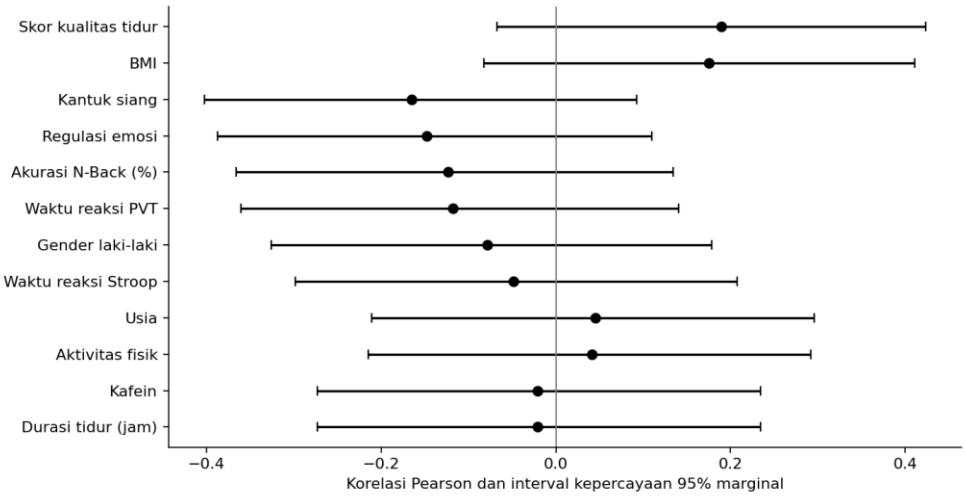
Korelasi maksimum 0,190 memiliki kuadrat sekitar 0,036, yang hanya menggambarkan asosiasi linier marginal dalam sampel dan bukan kontribusi prediktif tervalidasi. Interval yang lebar pada Tabel 5 memperlihatkan ketidakpresisian estimasi. Jendela waktu serta reliabilitas skor stres dan tes kognitif belum terdokumentasi, sehingga

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 56
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

ketidakselarasan konstruk merupakan penjelasan yang mungkin, bukan penyebab yang dapat dipastikan dari koefisien tersebut.

Literatur eksperimen mengenai tidur dan emosi [4] menelaah dampak manipulasi tidur pada luaran tertentu, sedangkan analisis ini memeriksa asosiasi antarindividu dalam satu tabel. Perbedaan pertanyaan, instrumen, dan desain dapat menghasilkan perbedaan temuan tanpa saling meniadakan. Demikian pula, konteks mediator psikologis pada studi Huang dan rekan [6] tidak terwakili secara lengkap oleh variabel fisik serta tugas kognitif di sini. Karena itu, korelasi yang rendah dibahas sebagai keterbatasan dataset untuk tugas prediksi tertentu, bukan sebagai penolakan terhadap teori hubungan tidur dan stres.



Gambar 4. Korelasi dan interval kepercayaan 95% marginal.

3.4 Evaluasi Bertingkat dan Pembandingan *Baseline*

Penerapan metode komparasi regresi pada Subbagian 2.4–2.6 menghasilkan evaluasi berpasangan pada partisi luar yang sama, sebagaimana dirangkum dalam Tabel 6. Dummy rata-rata memperoleh MAE 9,65 ± 1,33 dan RMSE 11,17 ± 1,25. Di antara model berbasis fitur, *RFE–Ridge* memiliki MAE terendah sebesar 9,69 ± 1,35, tetapi tetap lebih tinggi sekitar 0,04 poin daripada dummy. *Ridge* memperoleh MAE 9,76 ± 1,41, sedangkan *Linear Regression* mencapai 10,15 ± 1,80. Seluruh rata-rata R^2 berbasis fitur bernilai negatif, dari −0,470 pada *Linear Regression* sampai −0,235 pada *RFE–Ridge*. Dengan ruang *tuning* yang digunakan, belum terlihat nilai tambah prediktif yang konsisten terhadap *baseline* tanpa fitur.

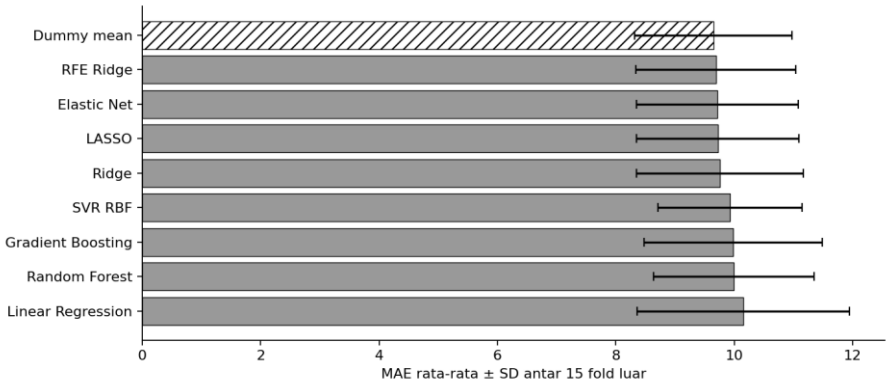
Tabel 6. Evaluasi bertingkat lima *fold* luar dengan tiga pengulangan

Pendekatan	MAE rata-rata ± SD	RMSE rata-rata ± SD	R^2 rata-rata ± SD
Dummy rata-rata	9,65 ± 1,33	11,17 ± 1,25	−0,214 ± 0,260
Linear Regression	10,15 ± 1,80	12,23 ± 1,55	−0,470 ± 0,379
Ridge	9,76 ± 1,41	11,32 ± 1,42	−0,255 ± 0,317
Elastic Net	9,72 ± 1,37	11,29 ± 1,35	−0,248 ± 0,308
SVR RBF	9,93 ± 1,21	11,45 ± 1,10	−0,300 ± 0,374
Random Forest	10,00 ± 1,35	11,62 ± 1,28	−0,325 ± 0,321
Gradient Boosting	9,98 ± 1,50	11,63 ± 1,44	−0,330 ± 0,383
LASSO	9,72 ± 1,37	11,31 ± 1,37	−0,251 ± 0,319
RFE–Ridge	9,69 ± 1,35	11,25 ± 1,29	−0,235 ± 0,285

Pada Gambar 5, rentang variasi antarpartisi saling tumpang tindih. Karena *fold* tidak independen dan sampel kecil, perbedaan peringkat deskriptif tidak cukup untuk menyatakan satu algoritma unggul. MAE *baseline* sekitar 9,65 setara sekitar 24,1% dari rentang skor teramati 40 poin; persentase ini hanya normalisasi deskriptif, bukan ambang manfaat klinis. Tidak tersedia *minimal clinically important difference* atau protokol skrining untuk mengubah kesalahan tersebut menjadi penilaian utilitas klinis.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 57
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 5. Perbandingan MAE pada 15 *fold* luar

R^2 negatif pada dummy juga memperlihatkan bahwa rata-rata data latih dapat berbeda dari rata-rata data uji pada sampel kecil. Karena itu, R^2 negatif tidak selalu berarti model lebih buruk daripada *baseline* yang memakai rata-rata data latih. Bukti bahwa model belum memberi keuntungan pada evaluasi utama berasal dari pembandingan langsung pada Tabel 6. Temuan ini lebih kuat daripada hanya menampilkan R^2 negatif tanpa menyertakan *baseline* yang benar-benar dihitung.

3.5 Dampak Regularisasi dan Seleksi Fitur

Tuning *Ridge* memilih $\alpha = 1000$ pada 13 dari 15 *fold* luar, $\alpha = 100$ pada satu *fold*, dan $\alpha = 10$ pada satu *fold*. Kecenderungan penyusutan kuat menunjukkan bahwa konfigurasi dengan koefisien kecil sering dipilih berdasarkan kinerja validasi dalam. Pada *LASSO*, seluruh koefisien menjadi nol pada 13 *fold*; dua *fold* lainnya mempertahankan lima dan delapan fitur. *Intercept* tetap ada ketika seluruh koefisien nol. Fenomena ini menjelaskan mengapa kinerja *LASSO* mendekati prediktor konstan, tetapi tidak membuktikan bahwa seluruh variabel tidak relevan pada populasi.

RFE memilih empat fitur pada tujuh *fold*, delapan fitur pada satu *fold*, dan 12 fitur pada tujuh *fold*. *RFE–Ridge* sedikit memperbaiki MAE relatif terhadap *Ridge* tanpa seleksi, tetapi belum mengungguli dummy. Kedua strategi seleksi fitur belum membuktikan nilai tambah terhadap *baseline* pada evaluasi ini. Jumlah fitur yang dipilih tidak ditafsirkan sebagai subset biomarker stabil karena pilihan berubah antarpartisi dan tidak ada validasi eksternal.

Model nonlinier juga belum memperbaiki rata-rata kesalahan dibanding dummy: MAE *SVR* sebesar 9,93, *Random Forest* 10,00, dan *Gradient Boosting* 9,98. Hasil ini melemahkan dugaan sederhana bahwa kegagalan semata-mata karena hanya model linier yang digunakan. Namun, kegagalan beberapa model nonlinier pada ruang pencarian terbatas tidak membuktikan ketiadaan pola nonlinier. Informasi target yang lemah, ukuran sampel, definisi fitur, dan heterogenitas pengukuran tetap dapat membatasi seluruh keluarga model secara bersamaan.

3.6 Analisis Sensitivitas Pembagian Tunggal

Tabel 7 melaporkan tiga pendekatan acuan pada pembagian 80:20 dengan seed 42. *Linear Regression* menghasilkan MAE 10,007, RMSE 11,704, dan R^2 −0,0179. Pada pembagian yang sama, dummy memiliki MAE 11,299 dan R^2 −0,1904; *Linear Regression* mengungguli dummy meskipun R^2 tetap negatif. Perbedaan ini mengilustrasikan pentingnya membedakan denominator R^2 dengan *baseline* rata-rata data latih.

Tabel 7. Sensitivitas Pembagian 80:20 Tanpa *Tuning* Pada Tiga Model Acuan

Pendekatan	MAE	RMSE	R^2
Dummy rata-rata	11,30	12,66	−0,1904
Linear Regression	10,01	11,70	−0,0179
Ridge	10,04	11,72	−0,0213

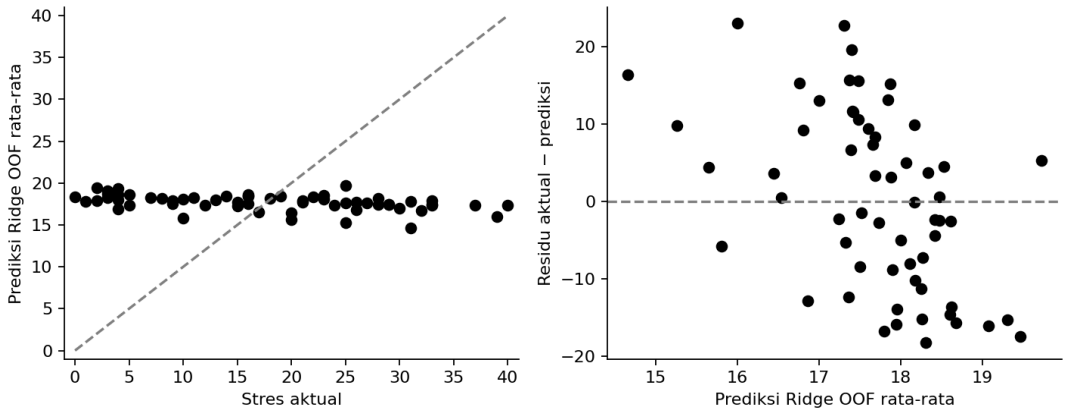
Ridge pada Tabel 7 menggunakan $\alpha = 1$, sedangkan *Linear Regression* dan dummy memakai konfigurasi dasar. Hasil pembagian tunggal tidak dapat dipadankan langsung dengan Tabel 6 yang mencakup *tuning* dan banyak partisi. Evaluasi utama menggunakan konfigurasi eksplisit dan keluaran per *fold* yang dapat diperiksa. Perubahan hasil antardesain ini menunjukkan sensitivitas evaluasi terhadap partisi dan regularisasi, bukan peningkatan klinis yang telah tervalidasi.

3.7 Diagnosis Residu dan Penyusutan Rentang Prediksi

Gambar 6 menunjukkan prediksi *Ridge* OOF rata-rata dan residunya. Target mencakup 0–40, tetapi prediksi rata-rata hanya berada pada 14,67–19,72. Model cenderung memprediksi terlalu tinggi pada skor aktual rendah dan terlalu rendah pada skor aktual tinggi. Rata-rata residu sebesar 0,166 mendekati nol, namun standar deviasinya 11,446. Dengan demikian, bias agregat yang kecil tidak berarti kesalahan individual kecil. Prediksi yang mengumpul dekat rata-rata konsisten dengan penyusutan koefisien kuat dan belum memadainya informasi untuk membedakan ekstrem target.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 58
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds



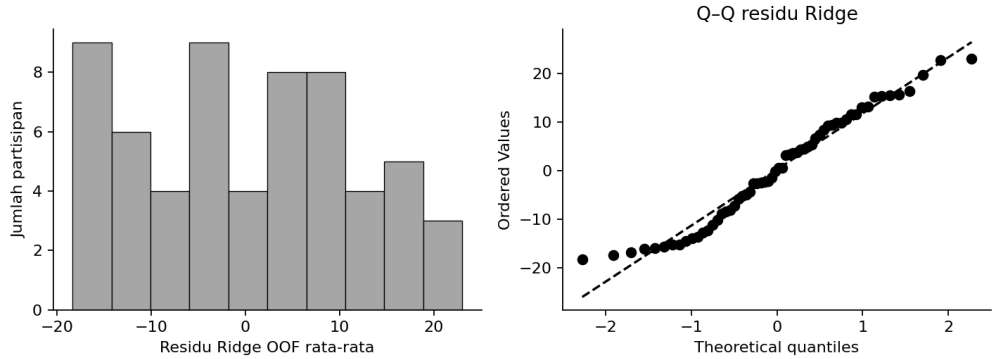
Gambar 6. Diagnosis *Ridge* dari rata-rata tiga prediksi OOF per observasi

Gambar 7 menambahkan histogram dan Q–Q residu. Shapiro–Wilk memberikan W = 0,960 dan p = 0,049 pada 60 residu agregat, yang secara eksploratif mengindikasikan penyimpangan dari normalitas. Nilai tersebut dekat ambang konvensional dan tidak digunakan untuk menetapkan mekanisme kegagalan. Normalitas residu juga bukan syarat perhitungan MAE maupun R^2. Korelasi Spearman antara |residu| dan prediksi sebesar −0,049 dengan p = 0,710 tidak menunjukkan pola monoton yang jelas, tetapi rentang prediksi yang sempit dan ketergantungan antarprediksi membatasi daya diagnosis heteroskedastisitas.

Selisih MAE *Ridge* dikurangi dummy dari prediksi OOF rata-rata per observasi adalah 0,107 poin. Bootstrap kondisional berpasangan menghasilkan interval persentil 95% sebesar −0,041 hingga 0,266, yang mencakup nol. Perhitungan ini menggunakan metrik atas prediksi rata-rata per orang sehingga tidak harus sama dengan selisih rata-rata metrik *fold* pada Tabel 6. Interval tersebut tidak mencakup ketidakpastian pelatihan ulang atau pemilihan model; hasilnya hanya mendukung kehati-hatian terhadap klaim selisih yang sangat kecil.

Standar deviasi residu *Ridge* sebesar 11,446 mendekati standar deviasi target 11,02; prediksi belum mengurangi variasi kesalahan secara memadai. *SVR*, *Random Forest*, dan *Gradient Boosting* juga tidak mengungguli dummy, sedangkan *LASSO* menolkan koefisien pada 13 dari 15 *fold*. Gabungan temuan ini tidak mendukung dugaan bahwa asumsi linier merupakan satu-satunya kendala. Namun, model nonlinier dalam ruang pencarian terbatas tidak mencakup semua bentuk fungsi; perbandingan ini tidak dapat memastikan dominasi *noise*, ketidakandalan label, atau ketidakselarasan konstruk.

Analisis residu belum dapat memisahkan apakah sumber kesalahan utama adalah *noise* pengukuran kognitif, ketidakandalan label stres, fitur yang tidak terukur, atau bentuk fungsi yang belum tertangkap. Untuk memisahkannya diperlukan pengukuran berulang yang memungkinkan estimasi reliabilitas, dokumentasi psikometri target, serta evaluasi *ablation* yang direncanakan pada sampel lebih besar. Penjelasan dalam bentuk kemungkinan mekanisme dibedakan dari hasil yang benar-benar teramati, yakni rentang prediksi sempit, kesalahan ekstrem, dan ketiadaan perbaikan terhadap *baseline*.



Gambar 7. Histogram dan Q–Q residu *Ridge* OOF rata-rata

3.8 Keterbatasan dan Arah Pengembangan

Ukuran sampel dan cakupan populasi membatasi generalisasi; resampling tidak menciptakan observasi baru. Instrumen, arah skor, satuan waktu reaksi, dan dosis kafein belum terdokumentasi lengkap. Evaluasi bersifat internal dan *post hoc*, dengan ruang *hyperparameter* terbatas. Interval korelasi, nilai p diagnostik, serta *bootstrap* kondisional bersifat eksploratif dan belum mencakup seluruh ketidakpastian. Tidak tersedia kohort eksternal atau pengujian prospektif.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 59
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

Feature importance internal *Random Forest* belum layak digunakan sebagai bukti biomarker ketika model tidak mengungguli *baseline*. Urutannya dapat dipengaruhi variasi sampel, pemilihan split, atau korelasi antarfitur. Evaluasi interpretabilitas berikutnya memerlukan performa yang tereplikasi, *permutation importance* pada data evaluasi, dan kestabilan antarresampling; hasilnya tetap tidak disamakan dengan efek kausal.

Studi berikutnya memerlukan instrumen stres tervalidasi dengan periode rujukan dan reliabilitas yang jelas, *diary* atau aktigrafi beberapa hari, serta dosis dan waktu konsumsi kafein. Variabilitas denyut jantung dari *ECG* atau *PPG*, aktivitas elektrodermal, akselerometri, dan konteks psikososial dapat diuji dengan sinkronisasi waktu serta pengendalian mutu. Kortisol saliva memerlukan protokol sirkadian tersendiri dan tidak diasumsikan sebagai label stres universal.

Pengukuran berulang memungkinkan *baseline* model campuran atau longitudinal sebelum pengujian *LSTM*. *LSTM* memerlukan urutan temporal nyata, jumlah subjek, dan durasi pengamatan yang memadai. Pemisahan antarindividu digunakan untuk menguji generalisasi ke pengguna baru, sedangkan evaluasi temporal menilai prediksi masa depan dengan mencegah overlap jendela. Prioritas pengembangan ialah keuntungan yang konsisten terhadap *baseline* dan validasi eksternal.

3.9 Pembahasan

Temuan utama penelitian ini adalah belum adanya keuntungan prediksi yang konsisten dari 12 indikator terhadap *baseline* rata-rata: MAE dummy sebesar 9,65, sedangkan *RFE–Ridge* sebesar 9,69 dan *Random Forest* sebesar 10,00 (Tabel 6). Selisih tersebut menunjukkan bahwa menambah kompleksitas pemodelan belum memperbaiki prediksi skor stres pada dataset ini. Pembahasan berikut membandingkan hasil tersebut dengan temuan penelitian terdahulu, dengan memperhatikan kesetaraan target, fitur, dan desain evaluasi. Tabel 8 merangkum arah perbandingannya.

Tabel 8. Perbandingan temuan penelitian ini dengan penelitian terdahulu

Studi	Temuan terdahulu	Temuan penelitian ini	Makna perbandingan
Larasati dkk. [20], 2024	Random Forest: akurasi 1,00 pada lima kelas stres.	Random Forest: MAE 10,00; R^2 −0,325.	Manfaat model tidak berpindah otomatis lintasdataset dan target.
Tutunji dkk. [9], 2023	Gabungan EMA–fisiologi: galat klasifikasi 29,87%.	Gabungan fitur belum mengungguli MAE dummy 9,65.	Konteks dan pengukuran berulang perlu diuji.
Huang dkk. [6], 2024	Korelasi stres–PSQI: r = 0,38; n = 8.178.	Korelasi kualitas tidur–stres: r = 0,190; n = 60.	Instrumen dan ketidakpastian membatasi penyetaraan.
Kinoshita dkk. [10], 2024	Variasi titik tengah tidur lebih besar pada stres tertinggi.	Korelasi durasi tidur–stres: r = −0,021.	Variabilitas tidur berbeda dari durasi tidur tunggal.
Ichwani dkk. [22], 2026	Pipeline: akurasi 78,21% pada klasifikasi Titanic.	Pipeline bertingkat belum memberi keuntungan terhadap dummy.	Pencegahan kebocoran tidak menjamin keberhasilan prediksi.

Larasati dkk. [20] melaporkan akurasi 0,99 untuk *Decision Tree* dan 1,00 untuk *Random Forest* dalam klasifikasi lima tingkat stres dari delapan indikator fisiologis dan tidur. Pada penelitian ini, *Random Forest* justru menghasilkan MAE 10,00 dan rata-rata R^2 −0,325, serta belum mengungguli dummy. Dengan demikian, keberhasilan algoritma berbasis pohon pada studi tersebut tidak terulang sebagai keuntungan regresi dalam dataset yang dianalisis. Angka akurasi tidak dapat dikonversi menjadi MAE atau dibandingkan sebagai peringkat kinerja lintasstudi karena target kategoris berbeda dari *Stress_Level* numerik, sementara fitur dan datasetnya juga berbeda. Perbandingan ini menunjukkan bahwa kelayakan algoritma bergantung pada informasi yang dibawa fitur dan definisi luaran, bukan pada nama algoritmanya saja.

Tutunji dkk. [9] memperoleh galat klasifikasi 29,87% ketika menggabungkan *EMA* dan data fisiologis, lebih rendah daripada *EMA* saja (33,45%) atau fisiologi saja (36,11%) pada model individual. Sebaliknya, gabungan indikator tidur, kognitif, dan variabel kontekstual pada penelitian ini belum menghasilkan MAE yang lebih rendah daripada prediksi rata-rata. Kesamaan penggunaan beberapa domain fitur dengan demikian tidak menghasilkan manfaat yang sama. Studi Tutunji dkk. mempunyai pengukuran berulang dan target konteks minggu ujian atau kontrol, sedangkan penelitian ini menggunakan rekaman tabular tanpa urutan waktu. Perbedaan tersebut membuat keselarasan waktu dan personalisasi layak diuji pada studi berikutnya; keduanya belum dapat ditetapkan sebagai penyebab kegagalan model saat ini.

Huang dkk. [6] menemukan korelasi stres dan skor PSQI sebesar 0,38 pada 8.178 mahasiswa; skor PSQI yang lebih tinggi menunjukkan kualitas tidur yang lebih buruk. Pada penelitian ini, korelasi *Sleep_Quality_Score* dengan *Stress_Level* lebih kecil, yaitu 0,190 dengan interval kepercayaan −0,067 hingga 0,423. Interval itu masih mencakup 0,38 sehingga hasil ini tidak membuktikan perbedaan kekuatan hubungan secara statistik. Selain itu, arah dan instrumen *Sleep_Quality_Score* belum terverifikasi, sehingga tanda positif tidak langsung dapat disetarakan dengan makna PSQI. Temuan pada dataset kecil ini menambah batas interpretasi, tanpa menyangkal hubungan tidur dan stres yang dilaporkan Huang dkk.

Kinoshita dkk. [10] menemukan variasi titik tengah tidur yang lebih besar pada kelompok stres tertinggi ($p < 0,001$) melalui pemantauan berulang selama delapan minggu. Sementara itu, durasi tidur tunggal pada penelitian ini berkorelasi −0,021 dengan stres, dan rata-rata skor stres kelompok tidur <7 jam justru 0,76 poin lebih rendah daripada kelompok 7–9 jam (Tabel 4). Hasil tersebut tidak memperlihatkan gradien sederhana durasi tidur–stres dalam sampel ini. Perbandingan ini mengarah pada kebutuhan menguji keteraturan dan perubahan tidur, bukan hanya jumlah jam tidur pada satu rekaman; perbedaan indikator dan unit analisis mencegah penafsiran kedua temuan sebagai pertentangan langsung.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 60
This work is licensed under a Creative Commons Attribution 4.0 International License.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

Dari sisi evaluasi, Ichwani dkk. [22] melaporkan peningkatan akurasi klasifikasi Titanic dari 72,62% menjadi 78,21% dengan *pipeline* terintegrasi. Penelitian ini menerapkan pemisahan prapemrosesan dan *tuning* dari evaluasi luar, tetapi tetap memperoleh rata-rata R^2 negatif pada semua model berbasis fitur. Kesamaan prinsip pengendalian kebocoran karenanya tidak menjamin informasi prediktif tersedia pada setiap dataset. Hasil ini mendukung perhatian terhadap generalisasi pada tinjauan Vos dkk. [12] dan heterogenitas fitur serta label pada Razavi dkk. [13]. Tidak dilakukan eksperimen dengan kebocoran yang disengaja, sehingga selisih performa penelitian ini tidak digunakan untuk mengukur besarnya efek pencegahan kebocoran.

Prediksi *Ridge* yang terkonsentrasi pada 14,67–19,72 meskipun target mencakup 0–40 memperjelas bahwa kelemahan model terjadi pada pembedaan individu, terutama di ekstrem skor. Dibandingkan studi yang berhasil memanfaatkan pengukuran fisiologis atau longitudinal, kontribusi penelitian ini adalah bukti batas nilai tambah fitur pada dataset tertentu melalui *baseline* dan evaluasi bertingkat. Temuan Birrer dkk. [24] tentang reliabilitas pengukuran tidur serta Gardiner dkk. [25] tentang pengaruh dosis dan waktu kafein memberi dasar pengembangan mutu pengukuran. Penambahan sensor dan variabel tersebut tetap memerlukan pengujian terhadap *baseline*, penyelarasan waktu, dan validasi pada data independen.

4. KESIMPULAN

Studi eksploratif terhadap 60 observasi pada dataset *Sleep Deprivation and Cognitive Performance* menunjukkan bahwa indikator tidur, kognitif, psikologis, demografis, dan kebiasaan yang tersedia belum mendukung prediksi *Stress_Level* yang andal. Tujuan evaluasi nilai tambah fitur dijawab melalui pembandingan langsung dengan *baseline* rata-rata dalam validasi silang bertingkat, dengan pengodean, standardisasi, *tuning*, serta seleksi fitur dipelajari hanya pada data latih. Semua interval kepercayaan korelasi marginal melintasi nol; korelasi terbesar sebesar 0,190 ditemukan pada kualitas tidur. Semua pendekatan berbasis fitur memiliki rata-rata koefisien determinasi negatif, dari −0,470 hingga −0,235, sehingga kegagalan generalisasi tampak pada evaluasi internal ini. *Baseline* memperoleh MAE 9,65 ± 1,33, lebih rendah daripada *RFE–Ridge* 9,69 ± 1,35 dan *Ridge* 9,76 ± 1,41; perbedaan kecil tersebut tidak dijadikan bukti keunggulan statistik suatu algoritma. Tuning, regularisasi, dan seleksi fitur belum menghasilkan nilai tambah yang konsisten. Diagnosis *Ridge* memperlihatkan penyusutan rata-rata prediksi OOF ke rentang 14,67–19,72, meskipun target mencakup 0–40, sehingga ekstrem skor kurang terwakili. Kontribusi empiris penelitian adalah dokumentasi batas keterpelajaran pada dataset tertentu dan penegasan bahwa relevansi teoretis fitur tidak dengan sendirinya menjamin prediksi individual. Hasil negatif ini tidak menyangkal hubungan tidur dan stres pada populasi, serta tidak mendukung pemakaian model sebagai alat skrining atau diagnosis. Analisis residu belum membedakan *noise* pengukuran, ketidakandalan label, fitur yang tidak terukur, dan bentuk hubungan yang belum tertangkap; mekanisme tersebut tetap memerlukan pengujian independen. Penelitian lanjutan perlu merencanakan sampel serta protokol terstandar, menggunakan instrumen stres tervalidasi dengan periode rujukan yang jelas, mengumpulkan catatan tidur beberapa hari melalui *diary* atau aktigrafi, dan mencatat dosis serta waktu konsumsi kafein. Fitur fisiologis seperti variasi interval denyut jantung dari *ECG* atau *PPG*, aktivitas elektrodermal, dan akselerometri dapat diuji bersama konteks psikososial yang disinkronkan secara temporal. Kortisol saliva hanya layak dipertimbangkan dengan pengendalian waktu sirkadian dan validasi protokol. Jika tersedia deret waktu nyata per individu, model campuran perlu menjadi *baseline* sebelum *LSTM*; evaluasi harus memisahkan individu serta waktu sesuai tujuan generalisasi. Ukuran sampel kecil, keterbatasan dokumentasi instrumen, sifat *post hoc*, dan ruang *hyperparameter* terbatas tetap menjadi batas utama, sehingga validasi eksternal diperlukan sebelum penerapan.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: Ibnu Sina Maulana;
Metodologi: Ibnu Sina Maulana;
Perangkat Lunak: Ibnu Sina Maulana;
Validasi: Ibnu Sina Maulana;
Analisis Formal: Ibnu Sina Maulana;
Investigasi: Ibnu Sina Maulana;
Sumber Daya: Ibnu Sina Maulana;
Kurasi Data: Ibnu Sina Maulana;
Penulisan Draf Awal: Ibnu Sina Maulana;
Penulisan, Peninjauan, dan Penyuntingan: Ibnu Sina Maulana dan Hafiyyan Putra Pratama;
Visualisasi: Ibnu Sina Maulana;
Supervisi: Hafiyyan Putra Pratama.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 61
This work is licensed under a Creative Commons Attribution 4.0 International License.

5.2 Ketersediaan Data

Dataset sumber tersedia pada Kaggle (doi: 10.34740/KAGGLE/DSV/10569288) dan salinan OpenML ID 46754 versi 1 (https://www.openml.org/d/46754). Kode evaluasi, konfigurasi model, metrik setiap *fold*, dan prediksi *out-of-fold* tersedia dari penulis korespondensi berdasarkan permintaan.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (*Not applicable*) untuk pengumpulan data baru. Analisis sekunder ini memakai dataset publik tanpa perekrutan, intervensi, atau kontak dengan partisipan. Persetujuan etik studi primer belum dapat diverifikasi dari dokumentasi repositori.

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (*Not applicable*) untuk analisis sekunder ini. Tidak dilakukan pengumpulan persetujuan partisipan baru; dokumentasi persetujuan pada studi primer belum dapat diverifikasi dari repositori.

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan tidak memiliki kepentingan keuangan atau hubungan pribadi yang diketahui dapat memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

ChatGPT digunakan untuk membantu penyusunan dan penyuntingan bahasa, penelusuran awal literatur, serta penyusunan kode analisis dan visualisasi. Penulis bertanggung jawab atas pemeriksaan kode, hasil perhitungan, ketepatan rujukan, interpretasi temuan, dan isi akhir manuskrip. Penggunaan AI tidak menggantikan tanggung jawab ilmiah penulis.

REFERENCES

[1] S. Griggs, A. Harper, and R. L. Hickman, "A systematic review of sleep deprivation and neurobehavioral function in young adults," *Appl. Nurs. Res.*, vol. 63, p. 151552, Feb. 2022, doi: 10.1016/j.apnr.2021.151552.

[2] R. Crowley, E. Alderman, A.-H. Javadi, and J. Tamminen, "A systematic and meta-analytic review of the impact of sleep restriction on memory formation," *Neurosci. Biobehav. Rev.*, vol. 167, p. 105929, Dec. 2024, doi: 10.1016/j.neubiorev.2024.105929.

[3] S. Y. Choong *et al.*, "A meta-analytic investigation of the effect of sleep deprivation on inhibitory control," *Sleep Med. Rev.*, vol. 80, p. 102042, Apr. 2025, doi: 10.1016/j.smrv.2024.102042.

[4] C. A. Palmer *et al.*, "Sleep loss and emotion: A systematic review and meta-analysis of over 50 years of experimental research.," *Psychol. Bull.*, vol. 150, no. 4, pp. 440–463, Apr. 2024, doi: 10.1037/bul0000410.

[5] J. M. Dzierzewski *et al.*, "How much sleep do you need: A 10-year systematic review of National Sleep Foundation's sleep duration recommendations and examination of sex differences," *Sleep Health*, vol. 12, no. 4, pp. 599–611, Aug. 2026, doi: 10.1016/j.sleh.2026.05.011.

[6] Y. Huang, L. Yang, Y. Liu, and S. Zhang, "Effects of perceived stress on college students' sleep quality: a moderated chain mediation model," *BMC Psychol.*, vol. 12, no. 1, p. 476, Sep. 2024, doi: 10.1186/s40359-024-01976-3.

[7] C. Gardiner *et al.*, "The effect of caffeine on subsequent sleep: A systematic review and meta-analysis," *Sleep Med. Rev.*, vol. 69, p. 101764, Jun. 2023, doi: 10.1016/j.smrv.2023.101764.

[8] M. A. Alnawwar, M. I. Alraddadi, R. A. Algethmi, G. A. Salem, M. A. Salem, and A. A. Alharbi, "The Effect of Physical Activity on Sleep Quality and Sleep Disorder: A Systematic Review," *Cureus*, Aug. 2023, doi: 10.7759/cureus.43595.

[9] R. Tutunji *et al.*, "Detecting Prolonged Stress in Real Life Using Wearable Biosensors and Ecological Momentary Assessments: Naturalistic Experimental Study," *J. Med. Internet Res.*, vol. 25, p. e39995, Oct. 2023, doi: 10.2196/39995.

[10] S. Kinoshita *et al.*, "Assessment of Stress and Well-Being of Japanese Employees Using Wearable Devices for Sleep Monitoring Combined With Ecological Momentary Assessment: Pilot Observational Study," *JMIR Form. Res.*, vol. 8, p. e49396, May 2024, doi: 10.2196/49396.

[11] X. Yao *et al.*, "Where does stress happen? Ecological momentary assessment of daily stressors using a mobile phone app," *Ment. Health Sci.*, vol. 2, no. 2, p. e54, Jun. 2024, doi: 10.1002/mhs2.54.

[12] G. Vos, K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," *Int. J. Med. Inf.*, vol. 173, p. 105026, May 2023, doi: 10.1016/j.ijmedinf.2023.105026.

[13] M. Razavi *et al.*, "Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review," *JMIR Ment. Health*, vol. 11, p. e53714, Aug. 2024, doi: 10.2196/53714.

[14] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[15] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nat. Commun.*, vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 62
This work is licensed under a Creative Commons Attribution 4.0 International License.

ISSN 2745-7966 (Media Online)
Vol 6, No 1, October 2026, Page 51-63
DOI: 10.47065/bulletinds.v6i1.10128
https://ejurnal.seminar-id.com/index.php/bulletinds

[16] 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, p. 339, Aug. 2025, doi: 10.1007/s10462-025-11326-3.
[17] I. Tsamardinos, "Don't lose samples to estimation," *Patterns*, vol. 3, no. 12, p. 100612, Dec. 2022, doi: 10.1016/j.patter.2022.100612.
[18] G. S. Collins *et al.*, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, Apr. 2024, doi: 10.1136/bmj-2023-078378.
[19] Farouk Mohammed and Juma al Majid, "Sleep Deprivation & Cognitive Performance." Kaggle. doi: 10.34740/KAGGLE/DSV/10569288.
[20] S. S. A. Larasati, E. N. K. Dewi, B. H. Farhansyah, F. A. Bachtiar, and F. Pradana, "Penerapan Decision Tree dan Random Forest dalam Deteksi Tingkat Stres Manusia Berdasarkan Kondisi Tidur," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 11, no. 5, pp. 1043–1050, Oct. 2024, doi: 10.25126/jtiik.2024117993.
[21] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: with Applications in Python*. in Springer Texts in Statistics. Cham: Springer International Publishing, 2023. doi: 10.1007/978-3-031-38747-0.
[22] A. Ichwani, R. I. Kesuma, A. Setiawan, I. E. Wicaksono, and R. Hanifah, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning," *J. Tek. Inform. Jutif*, vol. 7, no. 1, pp. 391–410, Feb. 2026, doi: 10.52436/1.jutif.2026.7.1.5490.
[23] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed. Chapman and Hall/CRC, 2019. doi: 10.1201/9781315108230.
[24] V. Birrer, M. Elgendi, O. Lambercy, and C. Menon, "Evaluating reliability in wearable devices for sleep staging," *Npj Digit. Med.*, vol. 7, no. 1, p. 74, Mar. 2024, doi: 10.1038/s41746-024-01016-9.
[25] C. L. Gardiner *et al.*, "Dose and timing effects of caffeine on subsequent sleep: a randomized clinical crossover trial," *SLEEP*, vol. 48, no. 4, p. zsae230, Apr. 2025, doi: 10.1093/sleep/zsae230.

© 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 63
This work is licensed under a Creative Commons Attribution 4.0 International License.