Klasifikasi Risiko Gangguan Tidur Berbasis Random Forest dan Analisis Feature Importance

Tiara Sari[1], Siti Hafizah[2], Femmy[3], Nieha Sumaya[4], Muhammad Ifan Rifani Ihsan[5]
[1,2,3,4,5] Universitas Bina Sarana Informatika, Pontianak
Email: kuliah.tiarasari@gmail.com

Abstrak

Tidur berperan penting dalam menjaga kesehatan fisik dan mental, namun identifikasi risiko gangguan tidur masih menjadi tantangan karena dipengaruhi oleh berbagai faktor yang saling berkaitan. Penelitian ini bertujuan menerapkan algoritma *Random Forest* untuk mengklasifikasikan tingkat risiko gangguan tidur ke dalam kategori *Healthy, Mild, Moderate,* dan *Severe,* serta mengidentifikasi atribut yang paling berpengaruh melalui analisis *feature importance*. Berbeda dengan penelitian sebelumnya yang umumnya berfokus pada evaluasi performa klasifikasi, penelitian ini juga menganalisis *feature importance* untuk mengidentifikasi faktor-faktor yang paling berpengaruh sehingga meningkatkan interpretabilitas model. Penelitian menggunakan pendekatan kuantitatif dengan memanfaatkan *Sleep Health and Daily Performance Dataset* dari Kaggle yang terdiri atas 100.000 sampel. Sebanyak 12 atribut dipilih sebagai variabel prediktor dan satu atribut sebagai variabel target. Tahapan penelitian meliputi *data preprocessing* menggunakan *LabelEncoder* dan *StandardScaler*, pembagian data menjadi 80% data latih dan 20% data uji, pembangunan model Random Forest dengan parameter *n_estimators* = 100 dan *max_depth* = 10, serta evaluasi menggunakan metrik *accuracy, precision, recall, F1-score, confusion matrix,* dan *feature importance*. Hasil penelitian menunjukkan bahwa model memperoleh *accuracy* sebesar 81,72% dan *weighted precision* sebesar 81,27%. Analisis *feature importance* menunjukkan bahwa atribut *mental_health_condition, sleep_duration_hrs,* dan *stress_score* merupakan atribut yang paling berpengaruh dalam proses klasifikasi risiko gangguan tidur. Penelitian ini memberikan kontribusi berupa model klasifikasi yang tidak hanya memiliki performa yang baik, tetapi juga mampu mengidentifikasi faktor-faktor dominan yang memengaruhi risiko gangguan tidur melalui analisis *feature importance,* sehingga dapat mendukung pengembangan sistem deteksi dini berbasis *machine learning* di bidang kesehatan.

Kata Kunci: ***Random Forest, Gangguan Tidur, Klasifikasi, Machine Learning, Data Mining***

Abstract

Sleep plays an important role in maintaining physical and mental health; however, identifying the risk of sleep disorders remains challenging because it is influenced by multiple interrelated factors. This study aims to apply the Random Forest algorithm to classify sleep disorder risk into four categories—Healthy, Mild, Moderate, and Severe—and to identify the most influential attributes through feature importance analysis. Unlike previous studies that primarily focused on classification performance, this study also analyzes feature importance to identify the most influential factors affecting prediction results, thereby improving model interpretability. A quantitative approach was employed using the Sleep Health and Daily Performance Dataset obtained from Kaggle, consisting of 100,000 samples. Twelve attributes were selected as predictor variables and one attribute as the target variable. The research procedure included data preprocessing using LabelEncoder and StandardScaler, splitting the dataset into 80% training data and 20% testing data, developing a Random Forest model with n_estimators = 100 and max_depth = 10, and evaluating the model using accuracy, precision, recall, F1-score, confusion matrix, and feature importance. The results showed that the

proposed model achieved an accuracy of 81.72% and a weighted precision of 81.27%. Feature importance analysis identified mental_health_condition, sleep_duration_hrs, and stress_score as the most influential attributes in classifying sleep disorder risk. This study contributes by providing a classification model that not only demonstrates good predictive performance but also identifies the dominant factors influencing sleep disorder risk through feature importance analysis, making it valuable for supporting the development of machine learning-based early detection systems in healthcare.

Keywords: ***Random Forest, Sleep Disorder, Classification, Machine Learning, Data Mining***

PENDAHULUAN

Tidur merupakan salah satu kebutuhan fisiologis yang berperan penting dalam menjaga kesehatan fisik, mental, dan kualitas hidup seseorang. Kualitas tidur yang baik berkontribusi terhadap peningkatan fungsi kognitif, produktivitas, serta kesejahteraan, sedangkan kualitas tidur yang buruk dapat meningkatkan risiko berbagai gangguan kesehatan, seperti penurunan konsentrasi, gangguan emosional, depresi, penyakit kardiovaskular, hingga gangguan metabolisme (Wilanda et al., 2024). Kualitas tidur dipengaruhi oleh berbagai faktor yang saling berkaitan, antara lain usia, aktivitas fisik, pola hidup, durasi tidur, kondisi kesehatan mental, dan tingkat stres (Farizkia Dinda et al., 2023; Zhang et al., 2024; Huang & Id, 2024). Kompleksitas hubungan antarvariabel tersebut menyebabkan proses identifikasi risiko gangguan tidur menjadi sulit apabila hanya mempertimbangkan satu atau beberapa faktor secara terpisah. Oleh karena itu, diperlukan suatu pendekatan yang mampu menganalisis berbagai variabel secara bersamaan sehingga proses identifikasi risiko gangguan tidur dapat dilakukan secara lebih objektif, cepat, dan akurat.

Perkembangan teknologi informasi mendorong pemanfaatan data dalam jumlah besar untuk mendukung proses pengambilan keputusan pada berbagai bidang, termasuk bidang kesehatan. Salah satu pendekatan yang banyak digunakan adalah *Data Mining*, yaitu proses menemukan pola dan hubungan antarvariabel dari sekumpulan data sehingga menghasilkan informasi yang bermanfaat (Bianca Putri & Wahyu Wijayanto, 2022). Salah satu teknik dalam *Data Mining* adalah klasifikasi, yaitu proses pemberian label kelas terhadap suatu data berdasarkan karakteristik yang dimiliki (Srirahayu & Setya Pribadie, 2023). Dalam bidang kesehatan, teknik klasifikasi telah banyak diterapkan untuk membantu proses diagnosis penyakit, prediksi risiko kesehatan, serta mendukung pengembangan sistem pendukung keputusan berbasis kecerdasan buatan karena mampu mengolah data dalam jumlah besar secara lebih efektif dan konsisten (Riyandi et al., 2025; Wang et al., 2026).

Berbagai penelitian telah menerapkan algoritma *machine learning* untuk melakukan klasifikasi gangguan tidur maupun permasalahan kesehatan lainnya. (Riyandi et al., 2025) menerapkan algoritma *Random Forest* dan *HistGradientBoosting* untuk mengklasifikasikan gangguan tidur berdasarkan faktor gaya hidup dan kesehatan serta menunjukkan bahwa kedua algoritma mampu menghasilkan performa klasifikasi yang baik pada dataset berskala besar. (Zhang et al., 2024) menggunakan algoritma *Random Forest* untuk melakukan skrining *obstructive sleep apnea* berdasarkan fitur *heart rate variability* (HRV) dan menunjukkan bahwa algoritma tersebut mampu memberikan hasil klasifikasi yang baik pada kasus gangguan tidur. (Khasanah et al., 2025) membandingkan algoritma *Random Forest* dan *K-Nearest Neighbor* (KNN) dalam klasifikasi gangguan tidur dan memperoleh hasil bahwa *Random Forest* memberikan

performa yang lebih stabil pada data dengan karakteristik yang kompleks. Hasil penelitian tersebut menunjukkan bahwa algoritma *Random Forest* memiliki potensi yang baik untuk diterapkan pada berbagai permasalahan klasifikasi di bidang kesehatan. Temuan tersebut juga didukung oleh (Monowar et al., 2025). yang menunjukkan bahwa pendekatan *ensemble learning* mampu meningkatkan performa deteksi gangguan tidur pada dataset kesehatan melalui kombinasi beberapa model dan teknik penyeimbangan data. *Random Forest* merupakan salah satu algoritma *ensemble learning* yang membangun sejumlah *decision tree* menggunakan teknik *bootstrap sampling*, kemudian menggabungkan hasil prediksi setiap pohon melalui mekanisme *majority voting*. Pendekatan tersebut menghasilkan model yang lebih stabil, memiliki kemampuan generalisasi yang baik, serta mampu mengurangi risiko *overfitting* dibandingkan penggunaan satu decision *tree* (Ahmed Salman et al., 2024). Dibandingkan dengan algoritma klasifikasi lain seperti *K-Nearest Neighbor (KNN)* yang sensitif terhadap pemilihan nilai *k*, *Support Vector Machine (SVM)* yang memerlukan penentuan fungsi *kernel* secara tepat, serta *Decision Tree* yang lebih rentan mengalami *overfitting*, *Random Forest* memiliki keunggulan dalam menangani dataset berdimensi tinggi, mampu memodelkan hubungan nonlinier antarvariabel, dan relatif lebih stabil terhadap variasi data. Selain memiliki performa klasifikasi yang tinggi, *Random Forest* juga menyediakan analisis *feature importance* yang dapat digunakan untuk mengidentifikasi atribut-atribut yang paling berpengaruh terhadap hasil prediksi sehingga interpretasi model menjadi lebih mudah dilakukan. Karakteristik tersebut menjadi alasan utama pemilihan *Random Forest* pada penelitian ini, karena dataset yang digunakan memiliki jumlah sampel yang besar dengan berbagai atribut yang saling berkaitan, sehingga diperlukan algoritma yang mampu menghasilkan klasifikasi yang akurat sekaligus mudah diinterpretasikan (Ahmed Salman et al., 2024).

Meskipun berbagai penelitian telah menunjukkan bahwa *Random Forest* memiliki performa yang baik dalam klasifikasi gangguan tidur, sebagian besar penelitian masih berfokus pada peningkatan akurasi model atau perbandingan performa antar algoritma. Penelitian yang memanfaatkan *Sleep Health and Daily Performance Dataset* untuk mengidentifikasi atribut-atribut yang paling berkontribusi terhadap klasifikasi risiko gangguan tidur melalui analisis *feature importance* masih relatif terbatas. Oleh karena itu, penelitian ini tidak hanya mengevaluasi performa model klasifikasi menggunakan algoritma *Random Forest*, tetapi juga menganalisis kontribusi setiap atribut terhadap hasil prediksi. Kontribusi penelitian ini terletak pada penerapan algoritma *Random Forest* yang dipadukan dengan analisis *feature importance* untuk mengidentifikasi atribut yang paling berpengaruh terhadap klasifikasi risiko gangguan tidur. Dengan demikian, penelitian ini tidak hanya menghasilkan model prediksi dengan performa yang baik, tetapi juga memberikan interpretasi mengenai faktor-faktor yang berkontribusi terhadap tingkat risiko gangguan tidur.

Penelitian ini menggunakan *Sleep Health and Daily Performance Dataset* yang diperoleh dari Kaggle (Thalla, 2023). Dataset tersebut terdiri atas 100.000 sampel yang merepresentasikan karakteristik individu, aktivitas harian, kebiasaan tidur, dan kondisi psikologis. Dari keseluruhan atribut, dipilih 12 atribut sebagai variabel prediktor dan satu atribut target, yaitu *sleep_disorder_risk*, yang terdiri atas empat kategori, yaitu *Healthy*, *Mild*, *Moderate*, dan *Severe*.

Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan algoritma *Random Forest* untuk mengklasifikasikan tingkat risiko gangguan tidur menggunakan *Sleep Health and Daily Performance Dataset*. Kinerja model

dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*, sedangkan analisis *feature importance* digunakan untuk mengidentifikasi atribut yang paling berpengaruh terhadap proses klasifikasi. Hasil penelitian diharapkan dapat memberikan kontribusi terhadap penerapan *machine learning* di bidang kesehatan serta menjadi dasar bagi pengembangan sistem pendukung keputusan untuk deteksi dini risiko gangguan tidur.

METODE

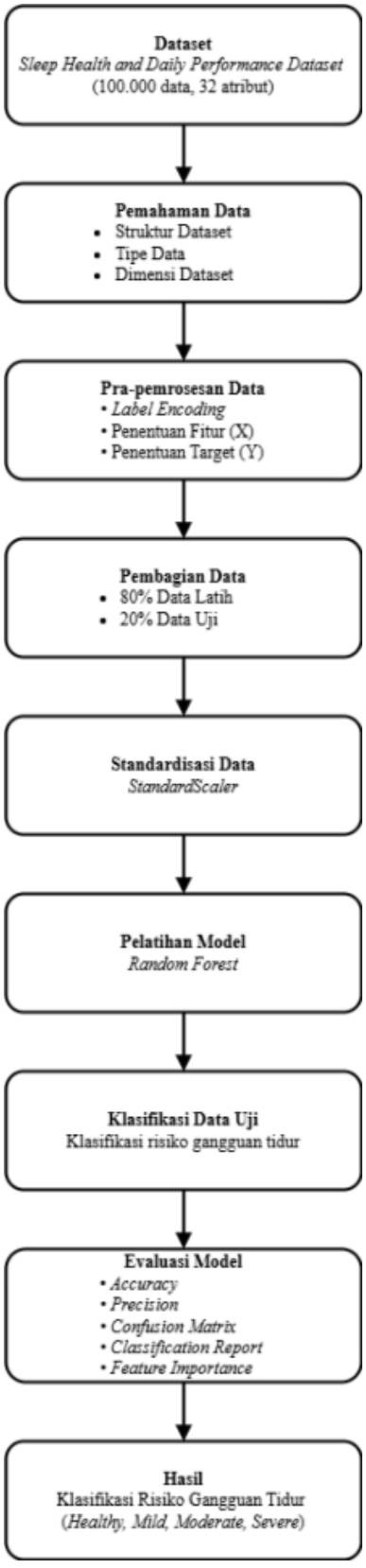
Penelitian ini menggunakan pendekatan kuantitatif dengan metode klasifikasi berbasis *machine learning* untuk mengklasifikasikan tingkat risiko gangguan tidur. Algoritma *Random Forest* dipilih sebagai model utama karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi, mengurangi risiko *overfitting* melalui mekanisme *ensemble learning*, serta menyediakan analisis *feature importance* yang dapat digunakan untuk mengidentifikasi atribut yang paling berpengaruh terhadap hasil klasifikasi.

Data dan Sumber Data

Dataset yang digunakan adalah *Sleep Health and Daily Performance Dataset* yang diperoleh secara publik melalui platform Kaggle (Thalla, 2023). Dataset terdiri atas 100.000 sampel dengan 32 atribut yang merepresentasikan karakteristik individu, pola tidur, aktivitas harian, dan kondisi psikologis. Dari keseluruhan atribut, dipilih 12 atribut sebagai variabel prediktor (*feature*) dan satu atribut target (*sleep_disorder_risk*) berdasarkan relevansinya terhadap tujuan penelitian.

Pra-pemrosesan Data

Tahapan penelitian secara sistematis disajikan pada Gambar 1.



Gambar 1. Alur Penelitian

Penelitian diawali dengan tahap pemahaman data (*data understanding*) untuk mengidentifikasi struktur dataset, tipe data, serta dimensi dataset. Tahap

berikutnya adalah pra-pemrosesan data (*data preprocessing*) yang meliputi transformasi atribut kategorikal menjadi data numerik menggunakan *LabelEncoder*, penentuan variabel prediktor (*feature/X*), serta penentuan variabel target (*Y*). Setelah proses pra-pemrosesan selesai, dataset dibagi menjadi 80% data latih dan 20% data uji menggunakan metode train-test split dengan random_state = 42. Penggunaan *random_state* bertujuan agar proses pembagian data dapat direproduksi pada penelitian selanjutnya. Selanjutnya dilakukan standardisasi data menggunakan *StandardScaler* untuk menyeragamkan skala atribut numerik sebelum proses pelatihan model. Meskipun *Random Forest* tidak mensyaratkan standardisasi, tahap ini tetap dilakukan untuk menjaga konsistensi tahapan pra-pemrosesan data. Metode *train-test split* dipilih karena penelitian ini berfokus pada pembangunan dan evaluasi awal model menggunakan dataset berukuran besar (100.000 sampel), sehingga pembagian data tersebut dinilai telah memadai untuk mengevaluasi performa model. Oleh karena itu, penelitian ini tidak menerapkan *k-fold cross validation*, yang menjadi salah satu keterbatasan penelitian dan direkomendasikan untuk diterapkan pada penelitian selanjutnya guna memperoleh evaluasi model yang lebih komprehensif..

Model Klasifikasi dan Parameter

Model klasifikasi dibangun menggunakan algoritma *Random Forest* dengan parameter n_estimators = 100, max_depth = 10, dan random_state = 42. Nilai n_estimators = 100 dipilih karena memberikan keseimbangan antara stabilitas model dan efisiensi waktu komputasi, sedangkan max_depth = 10 digunakan untuk membatasi kompleksitas pohon keputusan sehingga dapat mengurangi risiko *overfitting* tanpa mengurangi kemampuan model dalam mengenali pola data. Model kemudian dilatih menggunakan 80.000 sampel data latih untuk mempelajari hubungan antara variabel prediktor dengan variabel target sebelum digunakan untuk mengklasifikasikan data uji ke dalam empat kategori risiko gangguan tidur, yaitu *Healthy*, *Mild*, *Moderate*, dan *Severe*.

Evaluasi Model

Kinerja model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix* untuk mengukur kemampuan model dalam mengklasifikasikan tingkat risiko gangguan tidur. Selain itu, analisis *feature importance* dilakukan untuk mengidentifikasi kontribusi masing-masing atribut terhadap proses klasifikasi sehingga hasil prediksi dapat diinterpretasikan secara lebih komprehensif.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penerapan algoritma *Random Forest* dalam mengklasifikasikan tingkat risiko gangguan tidur menggunakan *Sleep Health and Daily Performance Dataset.* Pembahasan meliputi proses pengumpulan data, pra-pemrosesan, pembentukan model, evaluasi model, serta interpretasi hasil klasifikasi.

1. Tahap Pengumpulan Data

Penelitian menggunakan *Sleep Health and Daily Performance Dataset* yang diperoleh dari platform Kaggle (Thalla, 2023). Dataset terdiri atas 100.000 sampel dengan 32 atribut yang menggambarkan karakteristik individu, pola tidur, aktivitas harian, serta kondisi psikologis. Dari keseluruhan atribut dipilih 12 atribut sebagai variabel prediktor dan satu atribut target, yaitu *sleep_disorder_risk*, yang terdiri atas empat kategori, yaitu *Healthy*, *Mild*, *Moderate*, dan *Severe*.

Tabel 1. Ringkasan Dataset Penelitian

Keterangan	Nilai
Sumber data	Kaggle
Jumlah data	100.000
Jumlah atribut awal	32
Jumlah atribut yang digunakan	12
Variabel target	*sleep_disorder_risk*
Jumlah kelas	4

Berdasarkan Tabel 1, dataset memiliki jumlah data yang besar sehingga mampu merepresentasikan variasi karakteristik individu dan mendukung proses pembentukan model klasifikasi secara lebih optimal.

2. Tahap Pra-pemrosesan Data

Tahap pra-pemrosesan dilakukan sesuai dengan alur penelitian pada Gambar 1. Proses ini meliputi transformasi atribut kategorikal menggunakan *LabelEncoder*, penentuan variabel prediktor (*feature*) dan variabel target (*sleep_disorder_risk*), kemudian pembagian dataset menjadi 80.000 data latih (80%) dan 20.000 data uji (20%) menggunakan metode *train-test split* dengan random_state = 42. Setelah pembagian data dilakukan, atribut numerik distandardisasi menggunakan *StandardScaler* agar setiap atribut memiliki skala yang seragam sebelum proses pelatihan model.

3. Tahap Klasifikasi Menggunakan *Random Forest*

Model dibangun menggunakan algoritma *Random Forest* dengan parameter n_estimators = 100, max_depth = 10, dan random_state = 42. Parameter tersebut dipilih untuk menghasilkan model yang stabil sekaligus mengurangi kemungkinan terjadinya *overfitting*. Selanjutnya model dilatih menggunakan 80.000 data latih untuk mempelajari hubungan antara atribut prediktor dengan variabel target. Setelah proses pelatihan selesai, model digunakan untuk mengklasifikasikan 20.000 data uji ke dalam empat kategori risiko gangguan tidur.

4. Tahap Evaluasi Hasil

Ringkasan hasil evaluasi model disajikan pada Tabel 2.

Tabel 2. Hasil Evaluasi Model *Random Forest*

Parameter	Nilai
Accuracy	81,72%
Weighted Precision	81,27%
Jumlah data uji	20.000

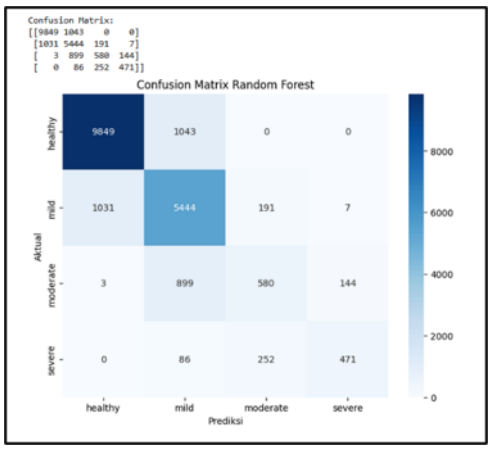
Model *Random Forest* memperoleh nilai *accuracy* sebesar 81,72% dan *weighted precision* sebesar 81,27%. Hasil tersebut menunjukkan bahwa model mampu mengidentifikasi pola hubungan antara karakteristik individu, kondisi psikologis, dan kebiasaan tidur dengan baik. Nilai tersebut menunjukkan bahwa model mampu mengenali pola hubungan antar atribut dengan baik sehingga sebagian besar data uji berhasil diklasifikasikan sesuai kelas sebenarnya.

Hasil penelitian ini sejalan dengan penelitian (Riyandi et al., 2025; Zhang et al., 2024) yang menunjukkan bahwa algoritma *Random Forest* mampu memberikan performa klasifikasi yang baik pada kasus gangguan tidur karena memanfaatkan mekanisme *ensemble learning* sehingga menghasilkan prediksi yang lebih stabil dibandingkan penggunaan satu *decision tree*. Meskipun demikian, performa model masih dipengaruhi oleh distribusi data pada setiap kelas sehingga kemampuan model dalam mengenali seluruh kategori belum sepenuhnya seimbang. Hasil *classification report* disajikan pada Tabel 3.

Tabel 3. Hasil *Classification Report*

Kelas Target	Precision	Recall	F1-Score	Support (Jumlah Sampel Uji)
Healthy	0,90	0,90	0,90	10.892
Mild	0,73	0,82	0,77	6.673
Moderate	0,57	0,36	0,44	1.626
Severe	0,76	0,58	0,66	809
Rata-rata (Weighted)	0,81	0,82	0,81	20.000

Berdasarkan Tabel 3, model menunjukkan performa terbaik pada kategori *Healthy* dengan nilai *F1-score* sebesar 0,90, sedangkan kategori *Moderate* memiliki nilai recall sebesar 0,36. Nilai tersebut menunjukkan bahwa model masih mengalami kesulitan mengenali seluruh sampel pada kategori *Moderate*, sehingga sebagian data diprediksi sebagai kategori lain yang memiliki karakteristik serupa. Distribusi hasil klasifikasi ditunjukkan pada Gambar 2.



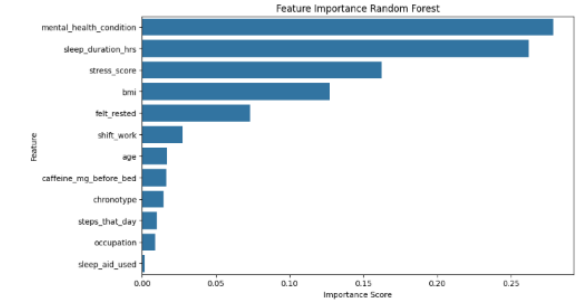
Gambar 2. *Confusion Matrix* Model *Random Forest*

Berdasarkan gambar 2, *confusion matrix* memperlihatkan bahwa sebagian besar data berhasil diprediksi sesuai kelas sebenarnya. Namun demikian, masih terdapat beberapa kesalahan klasifikasi terutama pada kategori *Moderate* dan *Severe*, yang menunjukkan bahwa model masih mengalami kesulitan membedakan kelas-kelas yang memiliki karakteristik serupa. Analisis *feature importance* ditampilkan pada Tabel 4.

Tabel 4. Nilai *Feature Importance*

Peringkat	Atribut Prediksi	Skor Kepentingan (Importance)	Persentase Kontribusi
1	mental_health_condition	0,27886	27,89%
2	sleep_duration_hrs	0,262048	26,20%
3	stress_score	0,162536	16,25%
4	bmi	0,127387	12,74%
5	felt_rested	0,073026	7,30%
6	shift_work	0,027516	2,75%
7	age	0,016984	1,70%
8	caffeine_mg_before_bed	0,016345	1,63%
9	chronotype	0,014524	1,45%
10	steps_that_day	0,010071	1,01%
11	occupation	0,009095	0,91%
12	sleep_aid_used	0,001607	0,16%

Berdasarkan Tabel 4, atribut *mental_health_condition* memiliki nilai *feature importance* tertinggi sebesar 27,89%, diikuti oleh *sleep_duration_hrs* sebesar 26,20% dan *stress_score* sebesar 16,25%. Hasil tersebut menunjukkan bahwa kondisi kesehatan mental, durasi tidur, dan tingkat stres merupakan faktor yang paling berpengaruh terhadap proses klasifikasi risiko gangguan tidur menggunakan algoritma *Random Forest*. Temuan tersebut sejalan dengan penelitian (Farizkia Dinda et al., 2023) yang menunjukkan adanya hubungan signifikan antara tingkat stres dan kualitas tidur. Selain itu, penelitian (Zhang et al., 2024) juga menunjukkan bahwa faktor-faktor yang berkaitan dengan kondisi tidur memiliki kontribusi penting dalam proses identifikasi gangguan tidur. Oleh karena itu, hasil penelitian ini memperkuat bahwa faktor psikologis dan kebiasaan tidur merupakan determinan utama dalam klasifikasi risiko gangguan tidur menggunakan algoritma *Random Forest*. Visualisasi nilai *feature importance* disajikan pada Gambar 3.



Gambar 3. Visualisasi *Feature Importance* Algoritma *Random Forest*

Gambar 3 memperlihatkan bahwa atribut *mental_health_condition*, *sleep_duration_hrs*, dan *stress_score* memiliki nilai *feature importance* yang lebih tinggi dibandingkan atribut lainnya. Hal tersebut menunjukkan bahwa ketiga atribut tersebut memberikan kontribusi terbesar dalam proses klasifikasi risiko gangguan tidur menggunakan algoritma *Random Forest*.

Pembahasan

Hasil pengujian algoritma *Random Forest* dalam klasifikasi tingkat risiko gangguan tidur menunjukkan nilai *accuracy* sebesar 81,72% dengan *weighted precision* sebesar 81,27%. Nilai tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali pola hubungan antara karakteristik individu, kondisi psikologis, dan kebiasaan tidur sehingga sebagian besar data uji berhasil diklasifikasikan dengan benar. Tingginya performa model didukung oleh karakteristik *Random Forest* yang membangun sejumlah *decision tree* melalui teknik *bootstrap sampling* kemudian menggabungkan hasil prediksi menggunakan mekanisme *majority voting*. Pendekatan tersebut mampu meningkatkan kemampuan generalisasi model, mengurangi risiko *overfitting*, serta menghasilkan prediksi yang lebih stabil dibandingkan penggunaan satu *decision tree* (Aini et al., 2024). Hasil penelitian ini juga sejalan dengan penelitian (Riyandi et al., 2025; Zhang et al., 2024) yang menunjukkan bahwa algoritma *Random Forest* mampu memberikan performa klasifikasi yang baik pada kasus gangguan tidur karena efektif dalam mempelajari hubungan antarvariabel pada data kesehatan.

Berdasarkan *classification report*, model menunjukkan performa terbaik pada kelas *Healthy* dengan nilai *precision*, *recall*, dan *F1-score* yang tinggi. Sebaliknya, performa terendah diperoleh pada kelas Moderate, terutama pada nilai *recall*. Kondisi tersebut menunjukkan bahwa model masih mengalami kesulitan mengenali seluruh data pada kategori *Moderate*, sehingga sebagian sampel diprediksi sebagai kelas lain. Fenomena tersebut juga terlihat pada *confusion matrix*, yang menunjukkan adanya kesalahan klasifikasi antara kelas *Moderate* dan *Mild*. Hal ini mengindikasikan bahwa kedua kelas memiliki karakteristik yang relatif mirip berdasarkan atribut yang digunakan sehingga batas antar-kelas menjadi kurang jelas. Temuan ini sejalan dengan penelitian (Ahmed Salman et al., 2024) yang menyatakan bahwa performa algoritma klasifikasi dipengaruhi oleh tingkat kemiripan karakteristik antar-kelas dan kualitas data yang digunakan dalam proses pelatihan.

Hasil analisis *feature importance* menunjukkan bahwa atribut *mental_health_condition*, *sleep_duration_hrs*, dan *stress_score* memiliki kontribusi terbesar terhadap proses klasifikasi. Temuan tersebut menunjukkan bahwa atribut yang merepresentasikan kesehatan mental, durasi tidur, dan tingkat stres memberikan kontribusi terbesar terhadap proses klasifikasi risiko gangguan tidur. Hasil ini selaras dengan penelitian (Farizkia Dinda et al., 2023) yang menunjukkan bahwa tingkat stres memiliki hubungan yang signifikan dengan kualitas tidur, serta (Wilanda et al., 2024) yang menjelaskan bahwa kualitas tidur dipengaruhi oleh berbagai faktor psikologis dan fisiologis. Selain itu, dominannya ketiga atribut tersebut menunjukkan bahwa faktor psikologis dan kebiasaan tidur memiliki pengaruh yang lebih besar dibandingkan atribut demografis, seperti usia dan pekerjaan, dalam proses klasifikasi risiko gangguan tidur menggunakan algoritma *Random Forest*.

Jika dibandingkan dengan penelitian terdahulu yang menerapkan algoritma Random Forest pada bidang kesehatan, hasil penelitian ini menunjukkan performa yang kompetitif dengan tingkat akurasi di atas 80%. Hasil tersebut sejalan dengan penelitian terdahulu yang

menunjukkan bahwa Random Forest mampu memberikan performa yang baik dibandingkan beberapa algoritma klasifikasi lain pada berbagai kasus klasifikasi gangguan tidur. Namun, penelitian ini tidak bertujuan untuk menentukan algoritma terbaik, melainkan berfokus pada evaluasi performa algoritma Random Forest serta analisis *feature importance* untuk mengidentifikasi atribut yang paling berpengaruh terhadap proses klasifikasi. Meskipun demikian, nilai recall yang relatif rendah pada kategori Moderate menunjukkan bahwa model masih mengalami kesulitan dalam membedakan karakteristik kelas tersebut dengan kelas lain yang memiliki pola serupa. Oleh karena itu, masih terdapat peluang untuk meningkatkan performa model melalui optimasi parameter maupun penerapan teknik pengolahan data yang lebih lanjut.

Penelitian ini memiliki beberapa keterbatasan, yaitu hanya menggunakan satu dataset publik sehingga belum sepenuhnya merepresentasikan kondisi populasi yang lebih luas. Selain itu, penelitian ini belum membandingkan performa *Random Forest* dengan algoritma klasifikasi lain seperti *XGBoost*, *LightGBM*, maupun *Support Vector Machine* (SVM). Penelitian selanjutnya diharapkan dapat menggunakan dataset yang lebih beragam, menerapkan teknik optimasi *hyperparameter* maupun metode penyeimbangan data apabila diperlukan, serta melakukan analisis komparatif terhadap beberapa algoritma klasifikasi untuk memperoleh model dengan performa yang lebih optimal.

Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma *Random Forest* mampu mengklasifikasikan tingkat risiko gangguan tidur dengan performa yang baik dan memberikan informasi mengenai atribut-atribut yang paling berpengaruh melalui analisis *feature importance*. Temuan ini menunjukkan bahwa algoritma *Random Forest* berpotensi menjadi dasar dalam pengembangan sistem pendukung keputusan berbasis *machine learning* untuk membantu proses identifikasi dini risiko gangguan tidur secara lebih objektif dan efisien.

SIMPULAN (PENUTUP)

Kesimpulan

Penelitian ini berhasil menerapkan algoritma *Random Forest* untuk mengklasifikasikan tingkat risiko gangguan tidur menggunakan *Sleep Health and Daily Performance Dataset.* Berdasarkan hasil pengujian, model memperoleh nilai *accuracy* sebesar 81,72% dan *weighted precision* sebesar 81,27%, sehingga menunjukkan bahwa algoritma *Random Forest* mampu mengklasifikasikan tingkat risiko gangguan tidur dengan performa yang baik pada dataset yang digunakan. Analisis *feature importance* menunjukkan bahwa atribut *mental_health_condition, sleep_duration_hrs,* dan *stress_score* merupakan faktor yang memberikan kontribusi terbesar terhadap proses klasifikasi. Hasil tersebut menunjukkan bahwa kondisi kesehatan mental, durasi tidur, dan tingkat stres merupakan faktor dominan yang memengaruhi risiko gangguan tidur.

Penelitian ini memberikan kontribusi ilmiah melalui penerapan algoritma *Random Forest* yang dipadukan dengan analisis *feature importance* untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap klasifikasi risiko gangguan tidur. Selain menghasilkan model klasifikasi dengan performa yang baik, penelitian ini juga meningkatkan interpretabilitas model sehingga dapat memberikan informasi yang mendukung pengembangan sistem pendukung keputusan dan deteksi dini risiko gangguan tidur berbasis *machine learning*. Temuan ini diharapkan dapat menjadi referensi bagi pengembangan penelitian di bidang informatika kesehatan serta penerapan kecerdasan buatan dalam mendukung pengambilan keputusan di bidang Kesehatan.

Saran

Penelitian selanjutnya disarankan

menggunakan dataset yang lebih beragam agar model dapat diuji pada karakteristik populasi yang lebih luas. Selain itu, optimasi *hyperparameter* serta penerapan teknik penyeimbangan data, seperti SMOTE, dapat dilakukan untuk meningkatkan performa model, terutama pada kelas yang memiliki tingkat klasifikasi lebih rendah. Penelitian berikutnya juga disarankan melakukan analisis komparatif dengan algoritma lain, seperti *XGBoost*, *LightGBM*, dan *Support Vector Machine* (SVM), sehingga dapat diperoleh model klasifikasi dengan performa yang lebih optimal. Model yang dihasilkan pada penelitian ini juga berpotensi dikembangkan menjadi sistem pendukung keputusan atau aplikasi berbasis web maupun perangkat bergerak untuk mendukung proses skrining awal risiko gangguan tidur secara lebih praktis.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Bina Sarana Informatika atas dukungan yang diberikan selama pelaksanaan penelitian. Ucapan terima kasih juga disampaikan kepada dosen pembimbing atas arahan, masukan, dan bimbingan selama proses penyusunan penelitian. Selain itu, penulis menyampaikan apresiasi kepada penyedia *Sleep Health and Daily Performance Dataset* melalui platform Kaggle yang telah menyediakan dataset sehingga penelitian ini dapat dilaksanakan dengan baik.

DAFTAR PUSTAKA

Ahmed Salman, H., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/https://doi.org/10.58496/BJML/2024/007

Aini, N., Arif, M., Tri Agustin, I., & Binti Toyibah, Z. (2024). Implementasi Algoritma Random Forest untuk Klasifikasi Bidang MSIB di Prodi Pendidikan Informatika. *JURNAL INFORMATIKA*, *11*(1), 11–16. https://doi.org/https://doi.org/10.31294/inf.v11i1.20637

Bianca Putri, N., & Wahyu Wijayanto, A. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Komputika: Jurnal Sistem Komputer*, *11*(1), 59–66. https://doi.org/10.34010/komputika.v11i1.4350

Farizkia Dinda, R., Budi Yulianti, A., & Ary Lantika, U. (2023). Hubungan Tingkat Stress dan Kualitas Tidur Pada Mahasiswa Fakultas Kedokteran Tahap Akademik Universitas Islam Bandung. *Bandung Conference Series: Medical Science*, *3*(1), 927–931. https://doi.org/10.29313/bcsms.v3i1.6888

Khasanah, N., Uki Eka Saputri, D., Aziz, F., & Hidayat, T. (2025). Studi Perbandingan Algoritma Random Forest dan K-Nearest Neighbors (KNN) dalam Klasifikasi Gangguan Tidur. *Computer Science (CO-SCIENCE)*, *5*(1), 17–25. https://doi.org/https://doi.org/10.31294/coscience.v5i1.5522

Riyandi, A., Fernandez, S., Wijayanto, S., & Sukmadiningtyas. (2025). Classification Of Sleep Disorders Based On Lifestyle And Health Factors Using Machine Learning. *Journal of Artificial Intelligence and Software Engineering*, *5*(2), 783–789. https://doi.org/10.30811/jaise.v5i2.6983

Srirahayu, A., & Setya Pribadie, L. (2023). Review Paper Data Mining Klasifikasi Data Mining. *JURNAL ILMIAH INFORMATIKA GLOBAL*, *14*(1), 7–12. https://doi.org/10.36982/jiig.v13i2.2307

Thalla, M. K. (2023). *Sleep Health and Daily Performance Dataset*. Kaggle. https://www.kaggle.com/datasets/mohankrishnathalla/sleep-health-and-daily-performance-dataset

Wilanda, A., Shofi Mubarok, A., Suprayitno, E., Sumarni, S., & Imran, S. (2024). Pengaruh Gaya Hidup Sehat, Kualitas Tidur, dan Pola Makan terhadap Tingkat Kesehatan Mental pada Pekerja Kantoran di Jawa Barat. *Jurnal Multidisiplin West Science*, *3*(1), 69–77. https://doi.org/https://doi.org/10.58812/jmws.v3i01.946

Zhang, C., Yu, L., Li, L., Zeng, P., & Zhang, X. (2024). Screening for moderate to severe obstructive sleep apnea by using heart rate variability features based on random forest algorithm. *Sleep and Breathing*, *28*, 2521–2530. https://doi.org/10.1007/s11325-024-03151-9

Huang, A. A., & Id, S. Y. H. (2024). *Comparison of model feature importance statistics to identify covariates that contribute most to model accuracy in prediction of insomnia*. *865*, 1–12. https://doi.org/10.1371/journal.pone.0306359

Monowar, M. M., Nobel, S. M. N., Afroj, M., Hamid, M. A., Uddin, M. Z., Kabir, M. M., & Mridha, M. F. (2025). *Advanced sleep disorder detection using multi-layered ensemble learning and advanced data balancing techniques*.

Wang, H., Zhang, C., Chen, B., Xie, Y., & Tian, P. (2026). *Development , validation , and visualization of a machine learning-based predictive model for depression risk in sleep disorder patients*.