VOL. 3. NO. 02 MARET 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

OPTIMASI PARAMETER K-NEAREST NEIGHBOR DAN NAIVE BAYES UNTUK KLASIFIKASI GANGGUAN TIDUR BERDASARKAN FAKTOR KESEHATAN

Mario Akbar Fadillah[1], Martanto[2], Yudhitira Arie Wijaya[3], Dodi Solihudin[4].

Program Studi Teknik Informatika[14]
Program Studi Manajemen Informatika[2]
Program Studi Sistem Informasi[3]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
marioakbar027@gmail.com

(*) Corresponding Author : marioakbar027@gmail.com
Published : 30 Maret 2026

Abstract—*Sleep disorders are a global health problem that significantly impacts human physical and mental health and productivity. Early detection and classification of sleep disorders can be improved through the application of machine learning technology capable of recognizing complex patterns in health data, with model performance highly dependent on parameter selection and optimization. This study aims to optimize the parameters of the K-Nearest Neighbor (KNN) and Naive Bayes algorithms using Grid Search techniques to improve the accuracy of sleep disorder classification based on health factors. The research method is quantitative with a comparative experimental approach using the Sleep Health and Lifestyle Dataset from Kaggle, which contains individual data with variables such as sleep duration, stress levels, physical activity, and blood pressure. The research stages include data cleaning, normalization, encoding, data sharing, initial model training, and parameter optimization using Grid Search with k-Fold Cross Validation. Performance evaluation is carried out based on accuracy, precision, recall, and F1-score metrics. The results showed that parameter optimization successfully improved the performance of both algorithms. The Naive Bayes model with the best parameters (var_smoothing = 0.4977) achieved the highest accuracy of 0.825, while the KNN model stabilized at 0.80 with optimal parameters (n_neighbors = 7, metric = manhattan, weights = uniform). Thus, the optimized Naive Bayes model was the best because it demonstrated a balance between accuracy, computational efficiency, and stability of classification results between classes. This study demonstrates that the application of hyperparameter tuning effectively improves model performance and has the potential to be developed into an efficient and accurate health data-based sleep disorder prediction system.*

Keywords: *Machine Learning, Grid Search, K-Nearest Neighbor, Naive Bayes, Sleep Disorders.*

Abstrak—Gangguan tidur merupakan permasalahan kesehatan global yang berdampak signifikan terhadap kesehatan fisik, mental, dan produktivitas manusia. Deteksi dini dan klasifikasi gangguan tidur dapat ditingkatkan melalui penerapan teknologi *machine learning* yang mampu mengenali pola kompleks dalam data kesehatan, dengan performa model yang sangat bergantung pada pemilihan dan optimasi parameter. Penelitian ini bertujuan untuk mengoptimalkan parameter algoritma K-Nearest Neighbor (KNN) dan Naive Bayes menggunakan teknik *Grid Search* untuk meningkatkan akurasi klasifikasi gangguan tidur berdasarkan faktor kesehatan. Metode penelitian bersifat kuantitatif dengan pendekatan eksperimental komparatif menggunakan *Sleep Health and Lifestyle Dataset* dari *Kaggle*, yang berisi data individu dengan variabel seperti durasi tidur, tingkat stres, aktivitas fisik, dan tekanan darah. Tahapan penelitian meliputi data *cleaning*, normalisasi, *encoding*, pembagian data, pelatihan model awal, serta optimasi parameter menggunakan *Grid Search* dengan *k-Fold Cross Validation*. Evaluasi kinerja dilakukan berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa optimasi parameter berhasil meningkatkan performa kedua algoritma, di mana model Naive Bayes dengan parameter terbaik (*var_smoothing* = 0.4977) mencapai akurasi tertinggi sebesar 0,825, sedangkan KNN stabil pada akurasi 0,80 dengan parameter optimal (*n_neighbors* = 7, metric = *manhattan*, *weights* = *uniform*). Dengan demikian, Naive Bayes hasil optimasi menjadi model terbaik karena menunjukkan keseimbangan antara akurasi, efisiensi komputasi, dan stabilitas hasil klasifikasi antar kelas. Penelitian ini membuktikan bahwa penerapan *hyperparameter tuning* efektif meningkatkan kinerja model dan

berpotensi dikembangkan menjadi sistem prediksi gangguan tidur berbasis data kesehatan yang efisien dan akurat.

Kata Kunci : *Machine Learning, Grid Search,* K-Nearest Neighbor, Naive Bayes, Gangguan Tidur.

INTRODUCTION

Gangguan tidur merupakan masalah kesehatan global yang semakin umum terjadi dan memiliki dampak signifikan terhadap kualitas hidup, produktivitas kerja, dan ekonomi sosial masyarakat. Secara global, prevalensi insomnia terus meningkat dan menyebabkan masalah ekonomi dan kesehatan yang serius [1]. Gangguan tidur memengaruhi kondisi kesehatan mental seperti depresi dan kecemasan, serta kondisi kesehatan fisik seperti sistem kekebalan tubuh dan metabolisme [2]. Dampak sosial dan ekonomi dari gangguan tidur juga sangat besar, dengan kerugian produktivitas yang signifikan di lingkungan kerja [3], [4].

Teknologi prediktif berbasis *machine learning* menawarkan solusi potensial untuk menangani gangguan tidur secara otomatis dan akurat. Jumlah orang yang mengalami gangguan tidur tanpa diagnosis terus meningkat, dan banyak pasien tidak dapat mendiagnosis gangguan tidur hingga mencapai tahap komplikasi kronis, seperti hipertensi atau gangguan kognitif [5]. Namun, masih ada tantangan dalam mengembangkan model klasifikasi terbaik yang dapat bekerja dengan baik pada data kesehatan yang kompleks dan beragam.

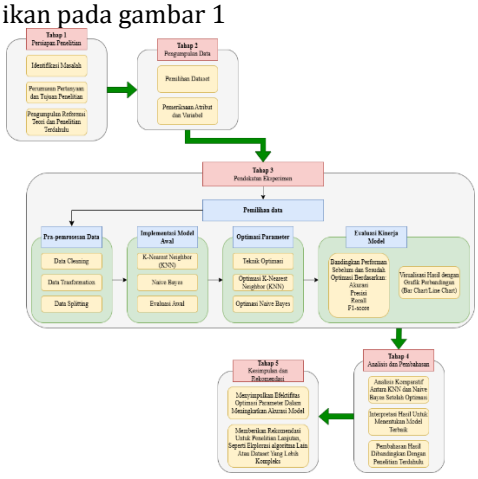
Hasil penelitian terdahulu menunjukkan bahwa beberapa algoritma *machine learning* telah digunakan untuk mengklasifikasikan gangguan tanpa menggunakan faktor fisiologis atau kesehatan. Metode berbasis deep learning, seperti CNN dan RNN, memberikan hasil yang dapat digunakan untuk mengklasifikasikan tahap tidur, tetapi memerlukan jumlah data yang besar dan daya komputasi yang tinggi [6]. Sebaliknya, model sederhana seperti K-Nearest Neighbor (KNN) dan Naive Bayes lebih efektif untuk dataset kecil atau terbatas, sehingga sering digunakan dalam penelitian yang berbasis faktor biologis dan kebiasaan tidur [7], [8].

Penelitian sebelumnya mengevaluasi 52 model *machine learning* dan menemukan bahwa algoritma seperti IBK (variasi dari KNN) dan Naive Bayes dapat memberikan kinerja tinggi dalam mengklasifikasikan gangguan tidur [7]. Hasil ini menunjukkan efektivitas kedua algoritma dalam mengklasifikasikan data medis, seperti penyakit ginjal kronis dan indeks perkembangan manusia [9], [10]. Meskipun banyak penelitian telah menggunakan KNN dan Naive Bayes dalam klasifikasi kesehatan, sebagian besar dari mereka masih menggunakan parameter default atau hanya sedikit mengeksplorasi ruang parameter secara sistematis. Hal ini seringkali menghasilkan model yang tidak ideal dan kurang memiliki generalisasi. Untuk mengatasi masalah ini, penelitian ini berfokus pada optimasi parameter menggunakan teknik *Grid Search* untuk menemukan kombinasi parameter optimal (K optimal dalam KNN, dan parameter smoothing dalam Naive Bayes), yang diharapkan dapat menghasilkan model klasifikasi gangguan tidur dengan kinerja yang lebih tinggi dan lebih andal.

Tujuan studi ini adalah untuk mengidentifikasi kesenjangan yang disebutkan di atas dengan mengoptimalkan parameter dalam algoritma KNN dan Naive Bayes untuk klasifikasi gangguan tidur berdasarkan faktor kesehatan. Diharapkan pendekatan ini dapat memberikan model yang lebih akurat dan efektif dalam memprediksi potensi gangguan tidur pada individu. Selain itu, penelitian ini bertujuan untuk mengevaluasi dampak variasi parameter terhadap kinerja model menggunakan analisis perbandingan dan data empiris [11], [12].

MATERIALS AND METHODS

Desain penelitian ini bersifat komparatif dan berorientasi pada replikasi, di mana eksperimen dilakukan untuk mengevaluasi dan membandingkan kinerja dua algoritma klasifikasi secara sistematis. Penelitian ini dilakukan menggunakan dataset *Sleep health and lifestyle* dari *Kaggle*. Langkah-langkah utama proses penelitian disajikan pada gambar 1



Gambar 1 Desain Penelitian

Diagram di atas menggambarkan seluruh proses penelitian yang dilakukan secara sistematis, mulai dari tahap persiapan hingga kesimpulan.

Pada Tahap 1 (Persiapan Penelitian), peneliti melakukan identifikasi masalah yang akan diteliti, kemudian merumuskan pertanyaan serta tujuan penelitian. Selain itu, dilakukan juga pengumpulan referensi dari penelitian terdahulu sebagai dasar teori dan pembanding. Tahap ini sangat penting karena menjadi fondasi utama agar penelitian memiliki arah yang jelas dan berbasis ilmiah.

Selanjutnya, pada Tahap 2 (Pengumpulan Data), peneliti memilih dataset yang relevan dengan permasalahan. Setelah itu dilakukan pemeriksaan terhadap atribut dan variabel dalam data untuk memastikan kualitas serta kesesuaiannya. Tahap ini menentukan kualitas hasil akhir, karena data yang baik akan menghasilkan model yang lebih akurat.

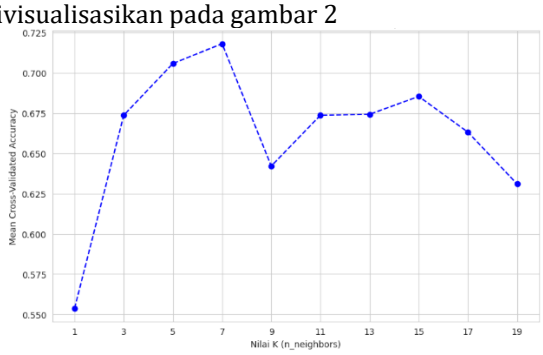
Masuk ke Tahap 3 (Pendekatan Eksperimen), proses menjadi lebih teknis dan terbagi ke beberapa bagian. Dimulai dari *pre-processing data* seperti data cleaning, transformasi, dan splitting. Kemudian dilakukan implementasi model awal menggunakan algoritma seperti K-Nearest Neighbor (KNN) dan Naïve Bayes. Setelah itu dilakukan optimasi parameter untuk meningkatkan performa model. Tahap ini diakhiri dengan evaluasi kinerja menggunakan metrik seperti akurasi, presisi, recall, dan F1-score, serta visualisasi hasil menggunakan grafik.

Pada Tahap 4 (Analisis dan Pembahasan), hasil eksperimen dianalisis dengan membandingkan performa model sebelum dan sesudah optimasi. Peneliti juga melakukan interpretasi untuk menentukan model terbaik serta membahas hasil tersebut dengan mengaitkannya pada penelitian sebelumnya.

Terakhir, Tahap 5 (Kesimpulan dan Rekomendasi) berisi rangkuman hasil penelitian, khususnya terkait efektivitas optimasi parameter dalam meningkatkan akurasi model. Selain itu, diberikan rekomendasi untuk penelitian selanjutnya, seperti eksplorasi algoritma lain atau penggunaan dataset yang lebih besar.

RESULTS AND DISCUSSION

Proses optimasi dalam algoritma K-Nearest Neighbor (KNN) dilakukan untuk mencari kombinasi parameter terbaik yang dapat meningkatkan kinerja model dibandingkan dengan hasil pelatihan awal. Dalam proses ini digunakan teknik *Grid Search Cross Validation* (*Grid Search CV*) untuk menguji berbagai variasi parameter seperti jumlah tetangga terdekat (*n_neighbors*), jenis metrik jarak (*metric*), dan metode pembobotan (*weights*). Tujuannya adalah memperoleh konfigurasi parameter yang menghasilkan akurasi rata-rata tertinggi pada proses *cross-validation*. Hasil hubungan antara nilai parameter K dan akurasi *cross-validation* divisualisasikan pada gambar 2



Gambar 2 Akurasi CV vs Nilai K untuk Optimasi K-Nearest Neighbor

Gambar 2 menunjukkan tren perubahan akurasi hasil *cross-validation* seiring variasi nilai parameter K pada algoritma KNN. Nampak bahwa peningkatan nilai *k* dari 1 sampai 7 sebanding dengan peningkatan akurasi, di mana puncak performa tercapai pada K = 7 dengan rata-rata akurasi sebesar 0.72. Setelah nilai itu, akurasi berfluktuasi dan cenderung menurun seiring bertambahnya jumlah tetangga, yang menunjukkan bahwa K yang terlalu tinggi dapat membuat model kehilangan kemampuan untuk mendeteksi pola minor dalam data. Dengan demikian, nilai K = 7 dipilih sebagai konfigurasi terbaik karena menawarkan keseimbangan antara kestabilan prediksi dan kerumitan model.

Setelah menentukan nilai parameter optimal, konfigurasi akhir dari hasil pencarian parameter terbaik ditunjukkan pada tabel 4.9.

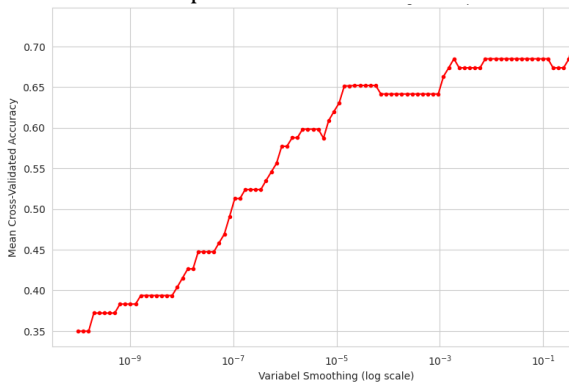
Tabel 1 Parameter Terbaik K-Nearest Neighbor

Parameter	Nilai Optimal
Metric	*manhattan*
n_neighbors	7
weights	*Uniform*
Akurasi Cross-Validation Terbaik	**0.7181**

Tabel 1 menampilkan hasil akhir konfigurasi parameter terbaik untuk model KNN berdasarkan hasil optimasi menggunakan *Grid Search CV*. Kombinasi *metric* = *'manhattan'*, *n_neighbors* = 7, dan *weights* = *'uniform'* memberikan performa terbaik dengan nilai akurasi *cross-validation* sebesar 0.7181. Pemilihan metrik *Manhattan*

distance dianggap paling sesuai untuk dataset penelitian ini karena mampu mengakomodasi perbedaan skala antar fitur dengan lebih baik dibandingkan metrik *Euclidean*. Hasil optimasi ini menunjukkan peningkatan efisiensi model dan menjadi dasar bagi pelatihan ulang model KNN dengan parameter teroptimasi pada tahap evaluasi berikutnya.

Proses optimasi pada algoritma Naive Bayes (GaussianNB) telah dilakukan untuk memperoleh nilai parameter *var_smoothing* terbaik yang memberikan performa klasifikasi tertinggi. Parameter *var_smoothing* berfungsi untuk menambahkan nilai kecil pada varians tiap fitur, sehingga mencegah terjadinya pembagian dengan nol dan meningkatkan stabilitas model. Proses pencarian nilai optimal dijalankan menggunakan teknik *Grid Search Cross Validation (Grid Search CV)* dengan rentang nilai var_smoothing antara 10^{-9} to 10^{0} pada skala logaritmik. Hasil pengujian terhadap berbagai nilai parameter tersebut divisualisasikan pada Gambar 3



Gambar 3 Akurasi CV vs Variabel Smoothing untuk Optimasi Naive Bayes

Gambar 3 memperlihatkan hubungan antara nilai *var_smoothing* dan akurasi hasil *cross-validation* model Naive Bayes. Diamati bahwa seiring peningkatan nilai *var_smoothing*, akurasi model juga mengalami kenaikan yang cukup konsisten hingga mencapai titik optimal. Nilai akurasi tertinggi tercatat pada *var_smoothing* sebesar 0.4977, dengan rata-rata akurasi *cross-validation* sebesar 0.7287. Setelah titik tersebut, performa model mulai mengalami fluktuasi ringan, yang menegaskan bahwa parameter ini memiliki pengaruh signifikan terhadap kemampuan generalisasi model. Hasil ini menunjukkan bahwa penambahan faktor smoothing yang sesuai mampu meningkatkan stabilitas dan akurasi model dalam menangani variasi data yang kompleks.

Setelah proses pencarian parameter terbaik dilakukan, konfigurasi hasil optimasi Naive Bayes dapat dilihat pada tabel 2 berikut.

Tabel 2 Parameter Terbaik Naive Bayes Hasil Optimasi

Parameter	Nilai Optimal
var_smoothing	0.497702356433211 37
***Akurasi Cross-Validation* Terbaik**	**0.7287**

Tabel 2 menunjukkan bahwa parameter terbaik untuk model Naive Bayes (GaussianNB) diperoleh dengan nilai *var_smoothing* sebesar 0.4977, nilai optimal ini menghasilkan akurasi *Cross-Validation* terbaik sebesar 72,87%. Hasil ini menegaskan bahwa penerapan teknik optimasi memberikan peningkatan signifikan dibandingkan dengan model awal yang hanya mencapai akurasi 0.65 (65%). Yang paling krusial, ketika diuji pada test set akhir menggunakan parameter optimal ini, Naive Bayes mencapai akurasi 82,5%, mengonfirmasi efektivitas optimasi parameter dalam menstabilkan model probabilistik pada dataset ini dan kemampuan generalisasinya.

CONCLUSION

Berdasarkan hasil eksperimen, penilaian model, dan proses optimasi parameter dengan *Grid Search*, studi ini menghasilkan wawasan yang jelas tentang kinerja awal model, efektivitas proses optimasi, serta perbandingan performa akhir antara algoritma K-Nearest Neighbor dan Naive Bayes. Kesimpulan dirumuskan untuk memberikan jawaban yang langsung dan terorganisir terhadap setiap pertanyaan penelitian, sebagai berikut:

Kinerja awal dari algoritma K-Nearest Neighbor dan Naive Bayes dalam mengklasifikasi gangguan tidur sebelum optimalisasi parameter menunjukkan bahwa kedua model mampu memberikan *baseline* yang memadai, tetapi belum mencapai performa terbaik. Pada langkah awal, model KNN mencatat akurasi 0.8000, sementara model Naive Bayes mencapai akurasi 0.6500, seperti yang terlihat pada evaluasi awal (merujuk pada Tabel 4.11 dalam dokumen). Nilai akurasi ini memberikan gambaran awal tentang kemampuan kedua model dalam mengenali pola gangguan tidur sebelum proses optimasi parameter dilakukan. Pada KNN, akurasi 0.8000 mencerminkan performa dasar yang cukup baik, namun masih dipengaruhi oleh pemilihan nilai *k* dan metrik jarak *default* yang belum tentu sesuai sepenuhnya dengan karakteristik dataset. Pada sisi lain, Naive Bayes dengan akurasi awal 0.6500 menunjukkan bahwa

asumsi independensi fitur dan parameter bawaan belum mampu memodelkan distribusi data dengan optimal. Oleh karena itu, kedua nilai akurasi ini menjadi pijakan untuk menilai sejauh mana optimasi parameter dapat meningkatkan kinerja masing-masing algoritma.

Penerapan metode *Grid Search* untuk mencari parameter optimal pada algoritma K-Nearest Neighbor dan Naive Bayes telah sukses mengidentifikasi kombinasi parameter terbaik yang dapat meningkatkan kinerja kedua model secara signifikan. *Grid Search* diterapkan secara sistematis dengan bantuan *k-Fold Cross Validation* untuk mengevaluasi seluruh kombinasi parameter pada kedua algoritma. Pada KNN, proses optimasi menghasilkan parameter terbaik berupa *n_neighbors* = 7, metrik jarak *Manhattan*, dan *weights* = *uniform*. Meskipun konfigurasi optimal ini memberikan stabilitas prediksi yang lebih baik, akurasi testing KNN tetap berada pada angka 0.8000, nilai yang sama dengan akurasi awal sebelum optimasi. Hal ini menunjukkan bahwa parameter *default* KNN sudah cukup mendekati konfigurasi optimal untuk dataset ini. Pada Naive Bayes, *Grid Search* mengevaluasi berbagai nilai *var_smoothing* dan menemukan nilai terbaik sebesar 0.4977. Parameter optimal ini meningkatkan kinerja model secara signifikan, dengan akurasi *testing* naik dari 0.6500 menjadi 0.8250. Dengan demikian, *Grid Search* terbukti lebih memberikan dampak peningkatan performa pada Naive Bayes dibandingkan KNN, namun tetap berperan penting dalam memastikan stabilitas dan konfigurasi parameter terbaik bagi kedua model.

Perbandingan kinerja model K-Nearest Neighbor dan Naive Bayes setelah dilakukan optimasi parameter menunjukkan bahwa Naive Bayes menjadi model terbaik dengan performa paling tinggi dan stabil. Setelah proses optimasi, kedua algoritma mengalami peningkatan performa, namun Naive Bayes menunjukkan peningkatan paling signifikan. Dengan *var_smoothing* optimal, Naive Bayes mencapai akurasi tertinggi sebesar 0.825 dan memperlihatkan stabilitas prediksi antar kelas yang lebih baik dibandingkan KNN. Sementara itu, KNN memperoleh akurasi stabil sebesar 0.80 menggunakan konfigurasi optimal yang ditemukan melalui *Grid Search*. Meskipun selisih akurasinya tidak terlalu besar, Naive Bayes unggul dalam hal efisiensi komputasi, kemampuan generalisasi, serta konsistensi hasil, sehingga lebih sesuai diterapkan pada dataset kesehatan berskala kecil seperti pada penelitian ini. Dengan demikian, Naive Bayes dapat disimpulkan sebagai model terbaik dalam klasifikasi gangguan tidur berdasarkan faktor kesehatan.

REFERENCE

[1] Iannella, P. A, and B. M, "Estimation of the global prevalence and burden of insomnia: a systematic literature review-based analysis," *Diagnostics*, vol. 15, no. 9, p. 1088, 2025, doi: 10.3390/diagnostics15091088.

[2] N. Seighali, A. Abdollahi, A. Shafiee, and M. J. Amini, "The global prevalence of depression, anxiety, and sleep disorder among patients coping with Post COVID-19 syndrome (long COVID): a systematic review and meta-analysis," *BMC Psychiatry*, vol. 24, p. 105, 2024, doi: 10.1186/s12888-023-05481-6.

[3] D. R. Glick *et al.*, "Economic Impact of Insufficient and Disturbed Sleep in the Workplace," *PharmacoEconomics*, vol. 41, no. 8, pp. 829–846, 2023, doi: 10.1007/s40273-023-01249-8.

[4] Alger, B. A, and B. T, "The social and economic cost of sleep disorders," *Sleep*, vol. 44, no. 8, 2021, doi: 10.1093/sleep/zsaa013.

[5] S. Ha, Y. Kim, and H. Hwang, "Predicting the Risk of Sleep Disorders Using a Machine-Learning-Based Questionnaire (SLEEPS): Development and Validation Study," *Journal of Medical Internet Research*, vol. 25, p. e46520, 2023, doi: 10.2196/46520.

[6] X. Xu, F. Cong, Y. Chen, and J. Chen, "Automatic Sleep Stage Classification Using Deep Learning: Signals, Representations and Performance," *Artificial Intelligence Review*, 2024, doi: 10.1007/s10462-024-10926-9.

[7] R. Alazaidah, R. Al-Shdefat, M. Alqurashi, F. Alzahrani, K. Abouelmehdi, and S. A. Shamim, "Potential of Machine Learning for Predicting Sleep Disorders," *Frontiers in Digital Health*, 2023.

[8] M. L. D. Araujo *et al.*, "Status and Opportunities of Machine Learning Applications in Obstructive Sleep Apnea: A Narrative Review," *Computational and Structural Biotechnology Journal*, vol. 28, pp. 167–174, 2025, doi: 10.1016/j.csbj.2025.04.033.

[9] V. Wulandari, W. J. Sari, Z. Alfian, L. Legito, and T. Arifianto, "Implementation of Naïve Bayes Classifier and K-Nearest Neighbor Algorithms for Chronic Kidney Disease Classification," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 710–718, 2024, doi: 10.57152/malcom.v4i2.1229.

[10] R. Anggara, Y. T. O. Mukhti, Y. Kurniawati, and D. Fitria, "Comparison of Naïve Bayes and K-Nearest Neighbors Methods in Classifying Human Development Index by Districts/City Indonesia in 2022," *UNP Journal of Statistics and Data Science*, vol. 2, no. 4, pp. 483–488, 2024, doi: 10.24036/ujsds/vol2-iss4/319.

[11] D. Ferreira-Santos, P. Amorim, T. S. Martins, M. Monteiro-Soares, and P. P. Rodrigues, "Enabling Early Obstructive Sleep Apnea Diagnosis With Machine Learning: Systematic Review," *Journal of Medical Internet Research*, vol. 24, no. 9, p. e39452, 2022, doi: 10.2196/39452.

[12] H. Almutairi, G. M. Hassan, and A. Datta, "Machine-Learning-Based-Approaches for Sleep Stage Classification Utilising a Combination of Physiological Signals: A Systematic Review," *Applied Sciences*, vol. 13, no. 24, p. 13280, 2023, doi: 10.3390/app132413280.