Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Predict Diagnosis of ME/CFS and Depression Using K-Nearest Neighbor Classification Method

Wisnu Ikhwansyah Saputra[1*], Alif Nur Fathlii Amarta[2], Elkin Rilvani[3]

[1,2,3]*Universitas Pelita Bangsa, Indonesia*
ikhwansyahwisnu@gmail.com[1*], alifamarta23@gmail.com[2]

Abstract

Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and depression are two medical conditions that often present with overlapping symptoms, making accurate diagnosis difficult, especially in early stages. This study aims to develop a predictive application using the K-Nearest Neighbor (K-NN) classification algorithm to differentiate between ME/CFS and depression based on clinical and lifestyle-related features. A dataset consisting of 1,000 records and 16 attributes was obtained from a public repository. Preprocessing steps included label encoding for categorical features, mean and mode imputation for missing values, and Min-Max normalization for numerical features. The model was trained using 80% of the data and evaluated on the remaining 20% using Manhattan distance and a k-value of 10. The application was developed with an interactive user interface, enabling predictions based on user input. The model achieved an overall accuracy of 88.5%, with excellent performance in detecting depression and ME/CFS, but moderate performance in identifying comorbid cases. The findings suggest that K-NN can be effectively utilized to support differential diagnosis in mental health, particularly for conditions with overlapping clinical symptoms. Future enhancements may include the incorporation of additional features and algorithmic improvements to address limitations in comorbid case detection.

Keywords: *Classification, Depression, Diagnosis Prediction, K-Nearest Neighbor, ME/CFS, Mental Health*

1. Introduction

ME/CFS (Myalgic Encephalomyelitis/Chronic Fatigue Syndrome) is a chronic, complex disorder marked by persistent fatigue, cognitive impairment, and post-exertional malaise. Diagnosis is often difficult due to the absence of specific biomarkers and considerable symptom overlap with mental health conditions—particularly depression [1].Machine learning, especially K-Nearest Neighbors (K-NN), has been increasingly employed in Indonesia to aid mental health classification and diagnosis. Nurdiansyah et al. applied K-NN and Random Forest to detect mental health conditions among Indonesian university students, with K-NN achieving approximately 90% accuracy on an 80:20 data split [2]. Another study by Pamungkas et al. implemented K-NN within a Case-Based Reasoning (CBR) system to diagnose mental disorders such as depression, anxiety, and stress. Their model achieved 84.62% accuracy, with precision and recall around 88 % and 85 %, respectively [3]. Ismail et al. developed an expert system diagnosing somatization disorders using K-NN with Euclidean distance, achieving an accuracy of 92.5 % [4]. Meanwhile, Anisa et al. used K-NN to classify student stress levels, recording 91.58% accuracy, with an F1-score of ~74 % [5].

However, despite these promising applications, there has been no prior study in Indonesia that combines K-NN for the simultaneous prediction of ME/CFS and depression. Given the symptomatic overlap between the two conditions, misdiagnosis is common. Developing an algorithm that can accurately distinguish ME/CFS from depression could significantly improve early intervention and patient care.
This research aims to utilize a labeled clinical dataset distinguishing ME/CFS and depression cases, Implement and tune a K-NN model (optimizing hyperparameters like k-value, distance metric, and feature scaling), Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score, Deploy the validated model as a user-friendly web or desktop application for healthcare practitioners. This study makes a novel contribution to Indonesian research by creating and locally validating a K-NN-based diagnostic tool for ME/CFS versus depression. It also supports clinical decision-making and differential diagnosis in a context where local data-driven tools remain scarce.

2. Literature Review

In recent years, machine learning techniques have gained increasing attention in the field of mental health diagnostics. Among them, the K-Nearest Neighbors (K-NN) algorithm stands out for its simplicity, ease of implementation, and relatively high accuracy when applied to structured datasets.

Pamungkas et al. [1] implemented a K-NN algorithm within a Case-Based Reasoning (CBR) system to assist in diagnosing mental health conditions such as stress, depression, and anxiety among Indonesian students. Using a cosine similarity measure and a k-value of 2, their model achieved an accuracy of 84.62%, with a precision of 88%, recall of 85%, and an F1-score of 84%.Similarly, Nurdiansyah et al. [2] compared the performance of K-NN and Random Forest algorithms in detecting mental health conditions among students. Their findings showed that K-NN achieved up to 90% accuracy on an 80:20 training-testing split. However, accuracy decreased to approximately 83% when the data split was adjusted to 70:30, suggesting the importance of proper data partitioning and validation. Rahma et al. [3] focused on text-based classification of mental illnesses by utilizing TF-IDF vectorization and K-NN with Levenshtein distance. Their web-based application, built using the Flask framework, demonstrated a high classification accuracy of 93% in identifying symptoms from user input data. Agustiyar applied attribute-weighted K-NN for heart disease prediction, improving classic K-NN results (79.87% vs 65.89%) by intergrating feature weighting on the model [6]. Rinaldy et al. [7] applied fuzzy K-NN to help decision support for stroke diagnosis at health centers in Indonesia, the model able to classify stroke risk (low, medium, high) with 61.1% accuracy. Although some studies focused on non-mental medical applications, such as Praningki & Budi (cervical cancer) and Wardhani et al. (liver disease), both showed that K-NN remains competitive in predictive classification [8][9].

Beyond algorithmic development, several studies have highlighted the prevalence of depression and stress among Indonesian populations. A survey conducted among junior nursing students reported that 32.1% experienced depression, while 66.7% exhibited varying levels of stress [4]. Another study at Universitas Gadjah Mada revealed a statistically significant correlation ($r = 0.597$, $p = 0.000$) between academic stress and depression tendencies during the COVID-19 pandemic transition period [5]. Sopwatun Anisa et al. [10] applied K-NN to classify the level of stress in university students and achieved 91.58% accuracy with 76.10% precision, 73.11% recall, and 74.17% f1-score.

Despite the promising results of K-NN in the mental health domain, none of the aforementioned studies explored the differential diagnosis between ME/CFS and depression, a challenge due to overlapping symptoms such as fatigue, cognitive impairment, and low motivation. Thus, there remains a critical research gap in distinguishing between these two conditions using machine learning approaches in the Indonesian context.

This research aims to address this gap by implementing a K-NN-based classification model that differentiates between ME/CFS and depression, validated using a structured clinical dataset. The findings are expected to provide a practical contribution to early diagnosis support systems and digital health innovation in Indonesia.

3. Research Methology

3.1. Dataset and Preproccesing

The dataset used in this study was obtaiuned from Kaggle [11] and it contains 1,000 records and 16 attributes collected from simulated clinical assessments related to ME/CFS (Myalgic Encephalomyelitis/Chronic Fatigue Syndrome) and depression. Each record contains both objective and subjective measurements, including fatigue scores, pain perception, sleep quality, stress level, and social activity indicators. The target variable is the diagnosis, initially consisting of three categories: “ME/CFS”, “Depression”, and “Comorbid/Both.

The dataset includes both numerical and categorical features. Numerical features include age, sleep_quality_index, brain_fog_level, physical_pain_score, stress_level, depression_phq9_score, fatigue_severity_scale_score, pem_duration_hours, and hours_of_sleep_per_night. Categorical attributes include gender, work_status, social_activity_level, exercise_frequency, and meditation_or_mindfulness. Categorical features were encoded using label encoding, and missing values were handled using mean imputation for numerical fields and mode imputation for categorical ones. To ensure that all features contributed equally to the distance calculation in the K-NN algorithm, Min-Max normalization was applied to the numerical features. The transformation was performed using the following formula:

$$X_{normal} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where is the original feature value, and represent the minimum and maximum values of the feature respectively. After preprocessing, the dataset was randomly divided into 80% training data and 20% testing data. This split ratio was selected to ensure sufficient data for model learning while retaining a reliable subset for unbiased evaluation. The class distribution was stratified to preserve the balance between ME/CFS and Depression classes in both training and testing sets.

The processed dataset was then used to train a K-Nearest Neighbors (K-NN) model using a value of and the Manhattan distance metric.

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i| \quad (2)$$

The Manhattan distance was selected due to its robustness in handling high-dimensional data and categorical encoding, where absolute differences are more informative than squared deviations. The distance between any two instances and was computed as:

Where is the number of features, and are feature values of the respective instances.This comprehensive preprocessing and configuration stage ensured that the data was clean, balanced, and optimally prepared for classification tasks using the K-NN algorithm.

3.2. K-NN Classification Method

The K-Nearest Neighbors (K-NN) algorithm was chosen for this study due to its simplicity, interpretability, and strong performance in structured classification tasks. K-NN is a non-parametric, instance-based learning algorithm that does not require an explicit training phase. Instead, it stores the entire training dataset and makes predictions by computing distances between a test instance and its k nearest neighbors from the training set. This characteristic is especially valuable in healthcare-related studies, where model transparency and low complexity are often preferred.

In this research, the K-NN model was implemented using the scikit-learn library in Python, utilizing a value of $k = 10$, which was

selected empirically based on common practice and preliminary trials. The Manhattan distance metric was used to compute the similarity between instances, as it provides a stable measure for both numerical and encoded categorical features. The Manhattan distance $d(p, q)$.between two instances p and q is calculated using the following formula:

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i| \tag{3}$$

Where n represents the number of features, and p_i, q_i are the feature values of the respective data instances.

All features were preprocessed and normalized using Min-Max scaling, ensuring that no single feature dominated the distance calculation. The training dataset—comprising 80% of the total data—was used to store the reference instances. The remaining 20% was used for testing, where the model calculated the distances between each test sample and all training samples, selected the 10 nearest neighbors, and assigned the majority label among them as the prediction. To maintain simplicity and clarity, a uniform weighting scheme was used, meaning that all neighbors contributed equally to the classification decision. While more advanced variants such as distance-weighted K-NN could offer slight performance improvements, the uniform approach provides a strong baseline and preserves model interpretability.

The entire model training and evaluation process was conducted using a hold-out validation strategy with a stratified split, ensuring that the class balance between ME/CFS and Depression was maintained across both the training and testing sets. Model performance was assessed based on standard evaluation metrics, which are detailed in the next section.

4. Result

This prediction is implemented through web-based application developed using pyrhon framework called streamlit. The following is the user interface of ME/CFS amd depression diagnosis prediction implemented through web-based application.

Fig. 1: Application Initial View

Figure 1 shows the initial view of the app. The interface are divided into three sections: title, about section and medical explanation. The title section shows the title of the app named “Prediksi Diagnosis ME/CFS dan Depression Dengan Metode k-NN”. The about section

explains that this application is designed to differentiate between ME/CFS and depression based on the various factors such as fatigue scores, pain perception, sleep quality, stress level, and social activity indicators that entered by user. In addition, there's a section explains this application is using dataset originates from kaggle to help predict diagnosis of ME/CFS or depression. The medical explanation section provides explanation of the medical glossary that used in the form to guide the user while entering form.

Fig. 2: Input Form Interface

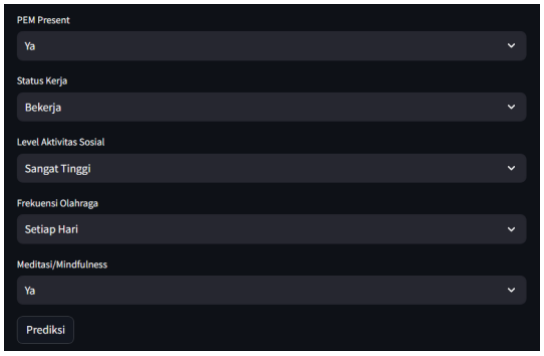
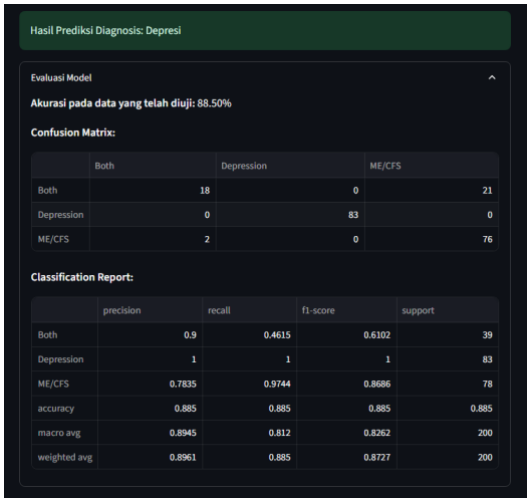


Fig. 3: Input Form Interface

Figure 2 and figure 3 is the input form interface, the form allows user to provide the relevant data to make a prediction of ME/CFS or depression. In figure 2 and figure 3, the user is asked to fill in demographic data in it's form, such as age and gender. Furthermore, there is assessment of health conditions and symptoms, such as sleep quality index, brain fog level, physical pain score, stress level, PHQ-9 depression score, fatigue level, duration of post-exertional malaise (PEM) symptoms in hours, average hours of sleep per night and the presence of PEM symptoms. Furthermore, there is lifestyle-related question including level of social activity, frequency of exercise and meditation or midfulness habits.

The input form of assessment of health conditions and symptoms are implemented through sliders except for presence of PEM symptoms. Presence of PEM symptoms and lifestyle-related question are implemented through dropdown selection. Once user has filled out the form, users can press predict button to see the result and it will trigger the model to calculate the data that provided by user using k-nearest neighbor.

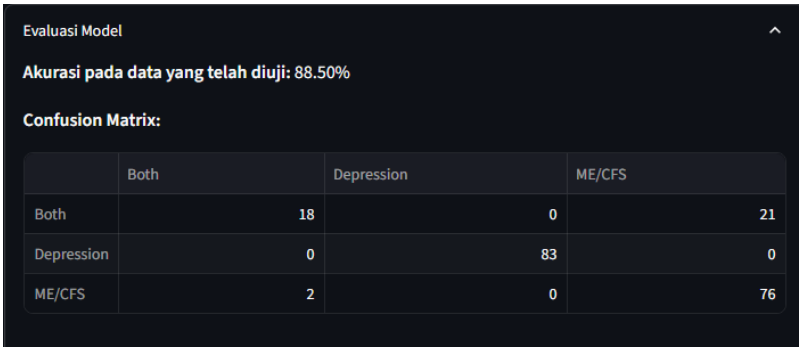


	Both	Depression	ME/CFS
Both	18	0	21
Depression	0	83	0
ME/CFS	2	0	76

	precision	recall	f1-score	support
Both	0.9	0.4615	0.6102	39
Depression	1	1	1	83
ME/CFS	0.7835	0.9744	0.8686	78
accuracy	0.885	0.885	0.885	0.885
macro avg	0.8945	0.812	0.8262	200
weighted avg	0.8961	0.885	0.8727	200

Fig. 4: Prediction Result and Performance Analysis

Figure 4 shows the predicted diagnosis based on user input. In this case the result of the prediction is depression, it means the model classified the nearest pattern based on the dataset using K-NN and it shows the result of prediction is depression. After the result, there's model evaluation.

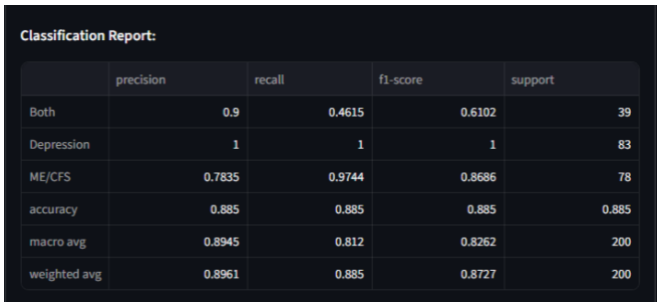


	Both	Depression	ME/CFS
Both	18	0	21
Depression	0	83	0
ME/CFS	2	0	76

Fig. 5: Model Evaluation

Figure 5 displays the result of model evaluation. The first part shows the accuracy on the tested data, it shows that the model is able to classify 88.5% of the data that have been tested. This accuracy is calculated based on the ratio between the number of correct prediction and the total number of test data. This high accuracy value indicates that the model able to performs well in recognizing symptom patterns that associated with depression, ME/CFS, and both.

The second part displays confusion matrix. the confusion matrix shows the breakdown number of correct and incorrect prediction for each diagnosis. "Both" was predicted correctly 18 times but there's 21 of them that predicted incorrectly and classified as "ME/CFS". "Depression" was predicted very well, there's 83 of them that was predicted correctly without any incorrect predictions. "ME/CFS" was predicted correctly 76 times, only 2 of them were predicted incorrectly and classifed as "Both". This means the model is really strong recognize "depression" and quite accurate for "ME/CFS" but is still not really optimal for recognizing "Both".



	precision	recall	f1-score	support
Both	0.9	0.4615	0.6102	39
Depression	1	1	1	83
ME/CFS	0.7835	0.9744	0.8686	78
accuracy	0.885	0.885	0.885	0.885
macro avg	0.8945	0.812	0.8262	200
weighted avg	0.8961	0.885	0.8727	200

Fig. 6: Classification Report

Figure 6 displays classification report,the classification report provides summary of classification model performance based on three main metrics precision, recoll, f1-score for each classes. Precision evaluate how accuracy the model is when predicting, recall evaluate the ability of the model to find the data that fit to the class, and f1 score is the harmonic mean between precision and recall (the higher the number the better) . "Depression" have 1.0 for every three main metrics. "ME/CFS" have 0.78 for precision, 0.97 for recall, 0.86 for f1-score. And "Both" have 0.90 for precision, 0.46 for recall, 0.61 for f1-score. The "depression" class had the best result with 1.00 for every three metrics and "both" class had the lowest recall with 0.4615, indicating that the model is often fails to detect both cases. (ME/CFS and depression).

5. Conclusion

This study developed a k-nearest neighbor model for distinguishing ME/CFS and depression with an interactive web-based pplication using simulation dataset from kaggle with 16 features such as fatigue level, sleep quality, and PHQ-9 depression score. After data goes through preprocessing (min-max normalization, label encoding, and missing value imputation), K-NN model with k = 10 and Manhattan distance metric achieved 88.5% accuracy. Performance analysis based on class showed excellent results for depression detection, high accuracy for ME/CFS, but for comorbid cases, the model have significant limitations in detecting the comorbid cases.

Although the model has the potential to reduce misdiagnosis due to symptom overlap, the main weakness is lack of capability to identify the combined condition of ME/CFS and depression. In the future, additional feature or algorithm optimization is needed to enhance the detection of more complex cases, as well as validation with real clinical data to ensure its capability in medical practice.

References

[1] A. Pamungkas, R. R. Isnanto, and D. M. K. Nugraheni, "Implementation of K-Nearest Neighbor in Case-Based Reasoning for Mental Health Diagnosis Systems," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 1109–1120, 2024.

[2] N. Nurdiansyah, F. S. Febriyan, Z. G. D. Amanta, D. A. Saputra, and W. M. Baihaqi, "Analisis Kesehatan Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Random Forest: Mental Health Analysis to Prevent Mental Disorders in Students Using The K-Nearest Neighbor (K-NN) Algorithm and Random Forest Algorithm," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 1–9, 2025.

[3] Y. Rahma, A. Prasetiadi, and M. Wibowo, "Identification Of Mental Ilness From Patient Diseases Using Knn And Levenshtein Distance Algorithm," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 5, pp. 1363–1372, 2022.

[4] A. Hendra, B. R. Heryanti, and A. L. Perdani, "Gambaran Tingkat Depresi, Kecemasan dan Stress pada Mahasiswa Junior Keperawatan di Indonesia," *Jurnal Keperawatan Komprehensif (Comprehensive Nursing Journal)*, vol. 6, no. 2, pp. 95–100, 2020.

[5] A. H. Azizah, S. Warsini, and K. P. Yuliandari, "Hubungan stres akademik dengan kecenderungan depresi mahasiswa ilmu keperawatan universitas gadjah mada pada masa transisi pandemi covid-19," *Jurnal Keperawatan Klinis Dan Komunitas (Clinical and Community Nursing Journal)*, vol. 7, no. 2, pp. 114–123, 2023.

[6] A. Agustiyar, "Prediksi Penyakit Jantung Menggunakan Attribute Weighting k-Nearest Neighbor," *InComTech : Jurnal Telekomunikasi dan Komputer*, vol. 13, no. 2, p. 145, Aug. 2023, doi: 10.22441/incomtech.v13i2.17883.

[7] Y. T. A. Rinaldy, A. A. Soebroto, and C. A. Setianto, "Sistem Pendukung Keputusan Diagnosis Penyakit Stroke menggunakan Metode Fuzzy K-Nearest Neighbor (FK-NN)(Studi Kasus Puskesmas Kendal Kerep Kota Malang)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 7, pp. 3149–3152, 2021.

[8] T. Praningki and I. Budi, "Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN," *Creative Information Technology Journal*, vol. 4, no. 2, pp. 83–93, 2018.

[9] A. K. Wardhani, L. Lakhmudien, A. N. Putri, and S. F. S. Ashour, "An Improved K-NN Algorithm and Bagging for Liver Disease Classification," *Telematika*, vol. 15, no. 2, pp. 100–107, 2022.

[10] S. Anisa, A. Komarudin, and E. Ramadhan, "Sistem Klasifikasi untuk Menentukan Tingkat Stress Mahasiswa Secara Umum Menggunakan Metode K-Nearest Neighbors," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 6, no. 3, pp. 568–578, 2024.

[11] Arshad Aliyev, "ME/CFS vs Depression Classification Dataset." Accessed: Aug. 05, 2025. [Online]. Available: *https://www.kaggle.com/datasets/storytellerman/mecfs-vs-depression-classification-dataset*