Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

Hybrid Two-Stage CatBoost and Multilayer Perceptron Model for Sleep Disorder Classification

Visal Ady Yanuar[1,*], Aripin[2]

[1] Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
[2] Faculty of Engineering, Dian Nuswantoro University, Semarang, Indonesia
Email: [1,*]visalady04@gmail.com, [2]arifin@dsn.dinus.ac.id
Correspondence Author Email: visalady04@gmail.com
Submitted: **14/05/2026**; Accepted: **30/06/2026**; Published**: 30/06/2026**

Abstract−Sleep disorders are a health problem that significantly impacts quality of life and potentially increases the risk of various chronic diseases. Conventional sleep disorder diagnosis generally requires expensive and complex examinations, so an alternative, non-invasive data-driven approach is needed. This study proposes a sleep disorder classification approach based on tabular medical data using a Hybrid Two-Stage architecture. The proposed approach integrates the CatBoost algorithm as a binary screening stage to distinguish healthy individuals from individuals with sleep disorders, and a Multilayer Perceptron (MLP) as a latent feature extractor, combined with CatBoost to classify sleep disorder subtypes, namely insomnia and sleep apnea. The datasets used were obtained from two public data sources and evaluated using a stratified k-fold cross-validation scheme. Class imbalance was addressed using the SMOTE-ENN technique, while hyperparameter optimization was applied as part of the model training pipeline. Performance evaluation was conducted using accuracy, Macro-F1, and Matthews Correlation Coefficient (MCC) metrics. Experimental results show that the Hybrid Two-Stage architecture achieves an accuracy of 94.1%, a Macro-F1 of 0.90, and an MCC of 0.88, and exhibits stable performance across a wide range of fold variations. These results demonstrate that the hybrid two-stage approach is effective in improving the performance of sleep disorder classification based on medical tabular data. The main contribution of this study is the development of a two-stage hybrid classification framework that explicitly separates healthy-disorder screening and sleep disorder subtype classification, while integrating SMOTE-ENN, Optuna-based hyperparameter optimization, and MLP-derived latent feature representation to improve classification stability on imbalanced medical tabular data.

Keywords: Sleep Disorders; Classification; Hybrid Two-Stage; CatBoost; Multilayer Perceptron

1. INTRODUCTION

Sleep is an active biological process that plays a critical role in maintaining human physiological and psychological balance. During sleep, the body performs essential restorative mechanisms, including hormonal regulation, memory consolidation, and cellular repair. Chronic sleep disorders have been widely associated with an increased risk of cardiovascular disease, metabolic disorders, and mental health conditions, positioning sleep disorders as a significant and growing global health concern [1], [2]. In clinical practice, sleep disorder diagnosis predominantly relies on polysomnography (PSG), a laboratory-based examination that comprehensively monitors physiological signals during sleep. Although PSG is considered the gold standard for sleep disorder diagnosis, it is associated with high costs, procedural complexity, and invasive settings. These limitations restrict its suitability for early screening and large-scale population-level applications, thereby motivating the development of alternative, non-invasive diagnostic approaches [3].

Advances in information technology and the increasing availability of digital health data have enabled new opportunities for non-invasive sleep disorder detection. Medical tabular data containing demographic information, lifestyle factors, and daily physiological indicators provide a practical foundation for building automated sleep disorder classification systems. In recent years, machine learning approaches have demonstrated considerable potential in identifying sleep disorder patterns from medical tabular datasets with competitive predictive performance [4], [5]. A wide range of machine learning techniques, including Support Vector Machines, Random Forests, and ensemble-based models, have been applied to sleep disorder classification tasks. Comprehensive reviews have highlighted that machine learning-based approaches are increasingly favored due to their flexibility and effectiveness in handling heterogeneous sleep-related data [6].

Despite achieving promising results, many studies still face a major challenge related to class imbalance, where healthy samples substantially outnumber samples representing specific sleep disorders. This imbalance often leads to biased models that favor the majority class and exhibit reduced sensitivity toward minority classes that are clinically more critical [7]. Beyond data imbalance, the complex non-linear relationships inherent in medical tabular data are often not adequately captured by a single modeling approach. Neural network-based models are capable of learning rich non-linear representations but tend to be unstable when directly applied to limited and imbalanced tabular datasets. Conversely, tree-based ensemble models typically provide higher stability and robustness but are less effective in extracting latent feature representations. Consequently, hybrid approaches that integrate the strengths of both paradigms have increasingly been proposed to improve classification performance and stability in sleep disorder detection [8], [9].

In addition to model architecture, hyperparameter selection is a crucial factor that significantly influences the performance and generalization capability of machine learning models. Manual hyperparameter tuning or conventional grid search methods are often inefficient and susceptible to subjective bias, particularly when dealing with large and complex search spaces. To address this limitation, Bayesian optimization-based hyperparameter tuning

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

has gained widespread adoption due to its ability to explore parameter spaces adaptively and efficiently. Optuna is one such framework that provides intelligent sampling strategies and automatic pruning mechanisms to accelerate the search for optimal hyperparameter configurations [10].

Several previous studies have investigated sleep disorder classification using structured health and lifestyle data. A summary of representative studies, including their methods and reported performance. One national-level study applied a Decision Tree algorithm to classify sleep disorders using lifestyle and health-related tabular data. The proposed model achieved an accuracy of 85.33%, with relatively balanced precision, recall, and F1-score values. The study identified daily steps, heart rate, sleep duration, and sleep quality as influential features for sleep disorder prediction. Although the reported performance was satisfactory, the reliance on a single classifier may limit prediction stability when facing data variability and class distribution shifts [11].

At the international level, a recent preprint study proposed the use of Large Language Models (LLMs) as an auxiliary system for sleep disorder classification on tabular data. In this work, LLMs were utilized to assist in designing and optimizing the classification pipeline through a prompting-based approach, while the final inference was still performed by conventional machine learning models. Experimental results showed that the LLM-assisted framework achieved an accuracy of 91.9% with an F1-score of 0.919. Despite its strong performance, the study positioned LLMs as a design-time support tool rather than an end-to-end classifier [12]. Another study focusing on structured health data evaluated several neural network architectures and hybrid models for sleep disorder classification. By combining neural networks with feature processing techniques, the proposed hybrid approach improved the representation of tabular data. The results indicated that the hybrid neural network achieved an accuracy of approximately 91%, demonstrating the potential of deep learning-based methods to capture non-linear relationships among tabular features. However, the proposed approach still employed a single-stage classification scheme [13]. In addition, a comparative study investigated multiple machine learning and deep learning models on lifestyle and health-related tabular datasets for sleep disorder prediction. The study reported that neural network-based models and other modern approaches were able to achieve accuracies of up to approximately 93%, depending on the model configuration and evaluation strategy. These findings highlight the effectiveness of advanced models for tabular data, while also emphasizing the sensitivity of performance to architectural design and parameter selection [14].

Based on these limitations, the research gap in previous studies lies in the lack of a classification framework that simultaneously addresses class imbalance, non-linear feature representation, hyperparameter optimization, and hierarchical sleep disorder prediction. Most existing approaches still classify all classes in a single stage, which may reduce model focus when distinguishing healthy samples from specific disorder subtypes. Therefore, this study proposes a Hybrid Two-Stage framework that separates the initial Healthy versus Disorder screening from the subsequent subtype classification stage. The main contribution of this study is the integration of CatBoost, Multilayer Perceptron-based latent feature extraction, SMOTE-ENN, and Optuna optimization within a unified two-stage pipeline to improve accuracy, class-balanced performance, and model robustness for sleep disorder classification using medical tabular data.

2. RESEARCH METHODOLOGY

This study proposes a hybrid two-stage classification framework for sleep disorder detection using structured (tabular) health and lifestyle data. The proposed framework is designed to address common challenges in medical tabular data, including class imbalance, limited sample size, and the presence of non-linear relationships among features. By separating the classification process into an initial screening stage and a subsequent subtype classification stage, the framework aims to improve prediction stability and generalization while maintaining model interpretability. The methodology emphasizes the integration of machine learning and deep learning techniques within a unified pipeline. Tree-based models are employed for their robustness on tabular data, while neural network–based representations are utilized to capture complex feature interactions. To ensure fair and reliable performance evaluation, all experiments are conducted using a stratified validation strategy. This section presents the dataset description, data preprocessing procedures, and the validation approach adopted in the experiments, which together form the foundation of the proposed classification framework. An overview of the proposed hybrid two-stage architecture is illustrated in Figure 1. The overall methodology is arranged to ensure that each stage of the proposed framework is executed systematically and independently within the validation process. This design is important to maintain the reliability of the experimental results, particularly because the dataset contains imbalanced class distributions and heterogeneous tabular features. Therefore, the proposed workflow combines preprocessing, imbalance handling, model optimization, two-stage classification, and evaluation into a single structured pipeline. Each component is designed to support the next stage so that the final classification result can be evaluated in a consistent and unbiased manner. In addition, the methodological design is structured to minimize possible data leakage during model development. The imbalance handling and hyperparameter optimization processes are applied only within the training portion of each fold, while the testing portion remains unseen until the evaluation stage. This procedure ensures that the reported performance reflects the model's ability to generalize to unseen data rather than memorizing the training samples. By applying this strategy, the proposed framework can provide a more reliable assessment of sleep disorder classification performance.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

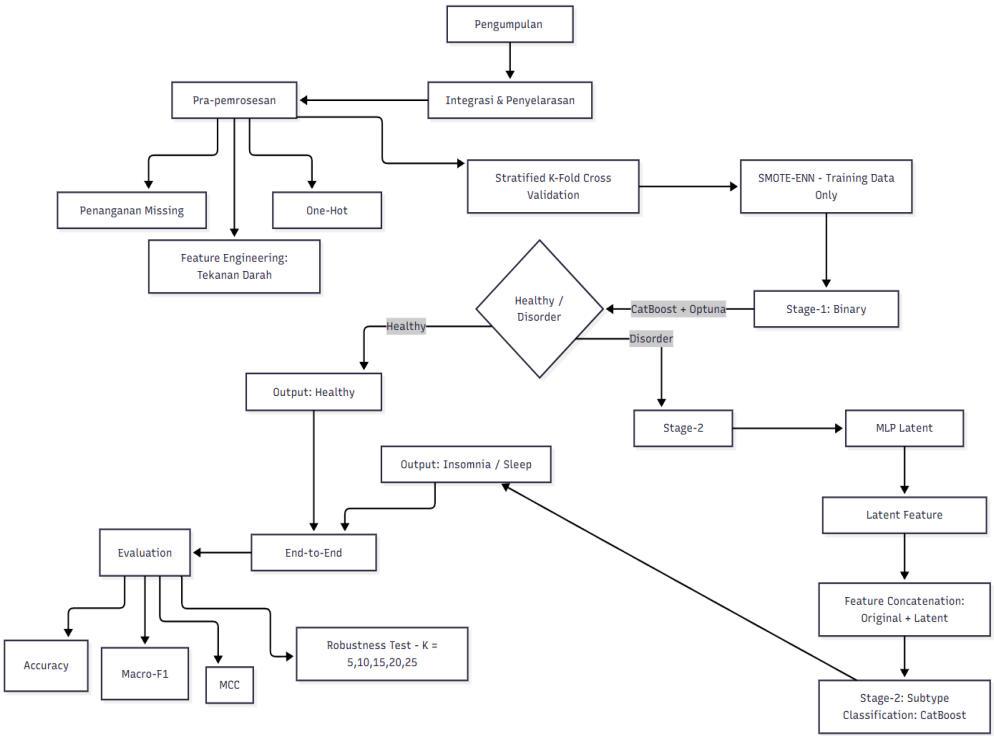

Figure 1. Workflow of the hybrid two-stage model for sleep disorder classification

2.1 Dataset

This study utilizes a tabular health and lifestyle dataset obtained from two public data sources, namely Kaggle and Figshare. The Kaggle dataset contains demographic information, physiological indicators, and lifestyle-related variables associated with individuals' sleep conditions. The Figshare dataset is used as an additional data source with a similar feature structure to complement the primary dataset. Both datasets are integrated to form a unified dataset used throughout the study. The integrated dataset consists of 560 samples with 13 attributes, including one target attribute representing the sleep condition of each individual. All attributes are represented in tabular form, making the dataset suitable for machine learning and deep learning approaches designed for structured data. The use of publicly available datasets also ensures the reproducibility of this study and supports non-invasive sleep disorder screening scenarios.

2.2 Data Preprocessing

The data preprocessing stage is performed to ensure data quality and consistency prior to model training. All integrated data are first aligned in terms of format and data types to ensure compatibility across attributes and to remove inconsistent values. The dataset was examined prior to preprocessing and no missing values were identified. Outlier removal was not performed, as extreme values in medical tabular data may represent valid clinical conditions rather than noise. In addition, the blood pressure attribute, which is originally represented as a single combined value, is separated into two numerical features: systolic blood pressure and diastolic blood pressure, in order to preserve physiological information more explicitly. Categorical features are then transformed into numerical representations so that they can be processed by machine learning and deep learning models. All preprocessing steps are applied consistently before data partitioning in the cross-validation scheme to prevent data leakage.

2.3 Stratified K-Fold Cross-Validation

All experiments in this study are conducted using Stratified K-Fold Cross-Validation, without employing a conventional train–test split. This strategy is selected to ensure that the class proportions in both training and testing data within each fold reflect the original class distribution of the dataset, which is inherently imbalanced. Formally, the dataset D is partitioned into K mutually exclusive subsets $D_1, D_2, \ldots, D_K$, such that the class distribution in each subset approximates that of the full dataset D. In the k- th iteration, subset Dk is used as the test set, while the remaining subsets are combined to form the training set. This process is repeated until each subset has been used exactly once as the test set. All modeling procedures are performed independently within each fold to obtain stable and unbiased performance estimates [15].

2.4 Handling Imbalanced Data Using SMOTE-ENN

The dataset used in this study exhibits an imbalanced class distribution, which may cause classifiers to be biased toward majority classes. To address this issue, SMOTE-ENN (Synthetic Minority Over-sampling Technique Edited

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

Nearest Neighbor) is employed to balance the training data while reducing noise. In the SMOTE phase, new synthetic samples are generated for the minority class by interpolating between a minority sample x_i and one of its k-nearest neighbors x_{nn}, as defined in (1).

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

As shown in (1), the synthetic sample x_{new} is generated through linear interpolation between the original minority instance and its nearest neighbor, where x_i denotes a minority class sample, x_{nn} represents one of its k-nearest neighbors, and $\lambda \in [0,1]$ is a random interpolation factor. After oversampling, the ENN procedure is applied to remove noisy samples. A data point x is removed if its class label differs from the majority class among its k -nearest neighbors. This process helps eliminate ambiguous samples and improves class boundary clarity. The SMOTE-ENN process is applied exclusively to the training data within each fold of the Stratified K-Fold Cross-Validation framework. This ensures that the test data preserves the original class distribution and prevents data leakage during evaluation[16].

2.5 Hyperparameter Optimization with Optuna

To obtain optimal model configurations, Optuna is employed as an automated hyperparameter optimization framework. Optuna formulates hyperparameter tuning as an optimization problem by maximizing an objective function defined on the training data. Formally, the optimization problem is defined as in (2).

$$\theta^* = \arg\max_{\theta \in \Theta} \quad f(\theta) \tag{2}$$

Where θ represents a set of model hyperparameters, Θ denotes the hyperparameter search space, $f(\theta)$ is the objective function, defined as the model performance (e.g., cross-validation score) on the training data. During optimization, Optuna iteratively samples candidate hyperparameters using Bayesian optimization strategies and evaluates them using the objective function. The best performing configuration θ^* is then selected for model training. Hyperparameter optimization is conducted only on the training data in each cross-validation fold, ensuring unbiased performance evaluation and improved generalization capability of the final model [10].

2.6 Stage-1: Binary Classification (Healthy vs Disorder)

Stage-1 of the proposed framework functions as an initial binary screening mechanism that distinguishes between Healthy and Disorder samples. This stage is designed to reduce the complexity of subsequent classification by filtering out normal cases at an early step, thereby allowing the following stage to focus exclusively on samples that potentially indicate sleep disorders. The binary classification model in Stage-1 is implemented using CatBoost, a gradient boosting algorithm based on an ensemble of decision trees [17]. In this model, each decision tree contributes a small score to the final prediction, and the overall prediction is obtained by aggregating the contributions of all trees in the ensemble. Mathematically, the prediction function of the Stage-1 classifier is defined as in (3).

$$\hat{y} = f_1(\mathrm{x}) = \sum_{m=1}^{M} h_m(\mathrm{x}) \tag{3}$$

Where x denotes the input feature vector, $h_m(\cdot)$ represents the m-th decision tree in the ensemble, M is the total number of decision trees, $\hat{y}$ is the aggregated prediction score obtained from all trees. The value of $\hat{y}$ represents the accumulated score resulting from the summation of the individual tree outputs, corresponding to the concept of additive scores in boosted decision trees. This aggregated score is then compared with a predefined decision threshold to determine the final binary class label. Formally, the classification decision is defined as in (4).

$$\hat{y}_{\text{class}} = \begin{cases} 1, & \text{if } f_1(\mathrm{x}) \geq \tau \\ 0, & \text{if } f_1(\mathrm{x}) < \tau \end{cases} \tag{4}$$

Where τ denotes the decision threshold, 1 corresponds to the Disorder class, 0 corresponds to the Healthy class. The CatBoost model in Stage-1 is trained using balanced training data obtained through the SMOTE-ENN procedure and employs the optimal hyperparameter configuration determined via Optuna. All training and optimization processes are conducted exclusively on the training data within each fold of the Stratified K-Fold Cross-Validation scheme to prevent data leakage. During inference, samples classified as Healthy are excluded from further processing, while samples classified as Disorder are forwarded to Stage-2 for subtype classification. This hierarchical two-stage strategy enables more focused learning, improves prediction stability, and enhances the overall generalization performance of the proposed sleep disorder classification framework.

2.7 Stage-2: Subtype Classification

Stage-2 is designed to perform sleep disorder subtype classification for samples identified as Disorder in Stage-1 by integrating latent feature learning and tree-based classification. Given an input feature vector $\mathrm{x} \in \mathbb{R}^{\mathrm{d}}$, a Multi-Layer Perceptron (MLP) is first employed to extract latent representations that capture non-linear interactions among the tabular features [18]. The latent feature learning process is defined as in (5).

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

$$z = g(x) \tag{5}$$

Where $g(\cdot)$ denotes the MLP mapping function and $z \in \mathbb{R}^k$ represents the learned latent feature vector of dimension k. The latent features z are then combined with the original input features through feature concatenation, resulting in an augmented feature vector, as defined in (6):

$$x' = [x; z] \tag{6}$$

where $[\cdot;\cdot]$ denotes the concatenation operator, and x' represents the augmented feature vector obtained by concatenating the original input features $x \in \mathbb{R}^d$, and the latent feature vector $z \in \mathbb{R}^k$. This fusion process integrates the original tabular information with high-level abstractions learned by the MLP, as illustrated in the proposed framework. The concatenated feature vector x' is subsequently used as input to a CatBoost-based multi-class classifier for subtype prediction. Similar to the binary classification stage, the CatBoost model aggregates the contributions of multiple decision trees to generate the final prediction score. Mathematically, the prediction function of the Stage-2 classifier is defined as in (7).

$$\hat{y} = f_2(x') = \sum_{m=1}^{M} h_m(x') \tag{7}$$

where $h_m(\cdot)$represents the m-th decision tree in the ensemble, M denotes the total number of trees, and $\hat{y}$ is the aggregated output score used to determine the predicted sleep disorder subtype. The final output corresponds to one of the predefined subtypes, such as Insomnia, Sleep Apnea, or other related sleep disorders.

2.8 Evaluation

Model performance is evaluated using Accuracy, Macro-F1 Score, and Matthews Correlation Coefficient (MCC) to provide a balanced assessment under imbalanced class conditions [19],[20], [21]. Accuracy measures the proportion of correctly predicted samples, Macro-F1 evaluates classification performance equally across all classes, and MCC is employed as a robust metric that considers all elements of the confusion matrix. These three metrics are used together to ensure that model performance is assessed not only from overall prediction accuracy, but also from class-balanced performance and prediction reliability. This combination of metrics is important because accuracy alone may be less representative for imbalanced datasets. The evaluation metrics are defined as in (8),(9),(10).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C} F\,1_i \tag{9}$$

$$\text{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

To provide a clearer overview of model performance, a confusion matrix is also utilized, summarizing the classification results in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These measures collectively describe correct and incorrect predictions across all classes and are particularly informative for imbalanced classification problems.

2.9 Robustness Check

To evaluate the stability and robustness of the proposed hybrid two-stage framework, additional experiments are conducted by varying the number of folds in the Stratified K-Fold Cross-Validation, namely 5, 10, 15, 20, and 25 folds. This analysis aims to ensure that the model performance does not depend on a specific validation configuration and remains consistent under different data partitioning schemes. For each fold configuration, the entire modeling pipeline including data preprocessing, imbalance handling, model training, and evaluation is executed independently. The robustness of the proposed approach is assessed by observing the consistency of Accuracy, Macro-F1 Score, and Matthews Correlation Coefficient (MCC) across different fold settings. Stable performance across these configurations indicates that the proposed method exhibits reliable generalization and is not sensitive to variations in the validation strategy.

3. RESULTS AND DISCUSSION

3.1 Optimization Analysis

Hyperparameter optimization is performed using Optuna for both CatBoost and MLP models. For the standalone CatBoost model, the optimization process converges to an objective value of approximately 0.86, with most performance gains achieved during the early trials. After convergence, subsequent trials provide only marginal improvements, indicating that Optuna is able to efficiently explore the hyperparameter search space and identify a stable and near-optimal configuration. As shown in Figure 2, this behavior suggests that CatBoost exhibits robust performance characteristics and benefits from hyperparameter tuning without requiring extensive fine-grained

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

adjustment. This indicates that Optuna contributes to improving model efficiency by reducing unnecessary trial exploration after reaching a stable configuration.

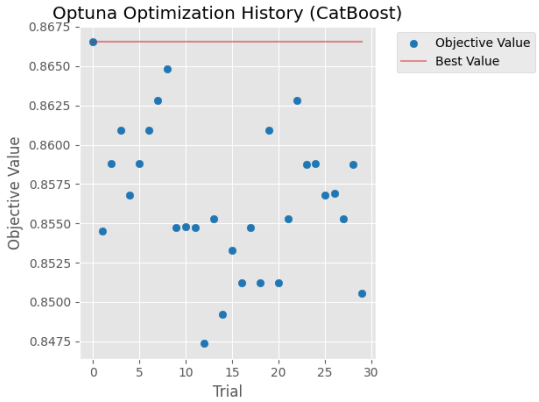

Figure 2. Optimization history of the CatBoost hyperparameters using Optuna

In the case of the standalone MLP model, Optuna improves the objective value to approximately 0.50, reflecting a noticeable improvement compared to the initial configuration. However, the optimized MLP performance remains substantially lower than that of CatBoost. As shown in Figure 3, this outcome indicates that while hyperparameter tuning contributes to performance enhancement, overall model effectiveness is also strongly influenced by architectural suitability. In particular, deep learning models applied directly to structured tabular data benefit less from hyperparameter optimization alone, highlighting the importance of combining architectural design with appropriate modeling strategies.

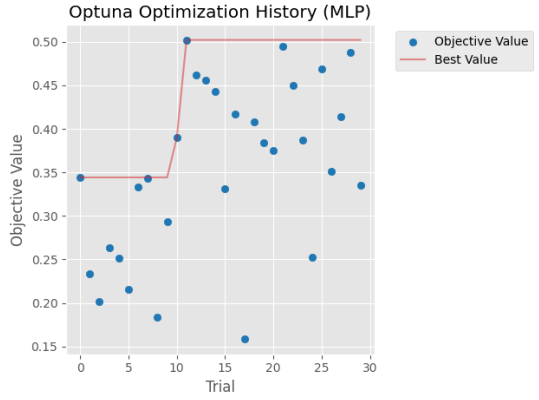

Figure 3. Optimization history of the MLP hyperparameters using Optuna

3.2 Learning Behavior Analysis

The learning behavior of the models is analyzed using training and validation loss curves to observe convergence patterns and generalization characteristics. For the standalone CatBoost model, the training loss decreases rapidly and approaches near-zero values, indicating strong learning capability on the training data. As shown in Figure 4, the validation loss initially decreases but begins to increase after approximately 40–50 iterations. This pattern suggests that, beyond a certain point, additional training iterations do not improve generalization and may lead to overfitting, even though the model continues to fit the training data well.

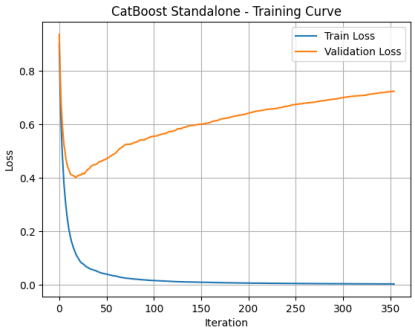

Figure 4. Training and validation loss curves of the standalone CatBoost model

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

The standalone MLP model shows a more gradual decrease in both training and validation loss throughout the training process. However, as shown in Figure 5, a persistent gap between the two curves remains, indicating that the model struggles to generalize effectively to unseen data. This observation is consistent with the relatively lower Macro-F1 and MCC values obtained by the MLP model, suggesting that learning non-linear patterns alone is insufficient when applied directly to structured tabular data. This result indicates that the standalone MLP requires additional feature support to improve its stability on tabular medical data.

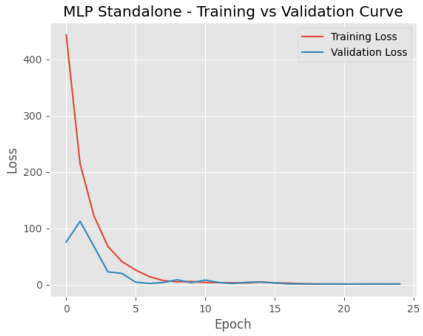

Figure 5. Training and validation loss curves of the standalone MLP model

In contrast, the hybrid two-stage model demonstrates more stable learning behavior in both Stage-1 (binary screening) and Stage-2 (subtype classification). Although a gap between training and validation loss is still observed, as shown in Figure 6, the validation loss converges and remains stable without significant divergence. This behavior indicates improved generalization performance, highlighting the benefit of combining latent feature learning through MLP with robust tree-based classification using CatBoost.

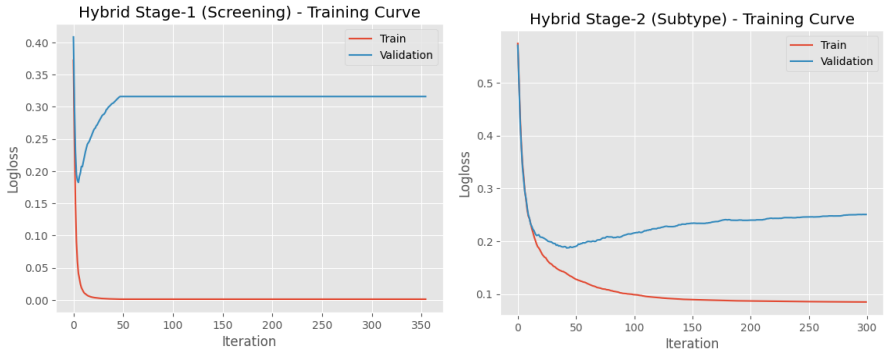

Figure 6. Training and validation loss curves of the Stage-1 and Stage-2 Hybrid model

3.3 Per-Fold Performance Analysis

The per-fold Macro-F1 results, as summarized in Table 2, reveal clear differences in both performance stability and effectiveness among the evaluated models. The standalone CatBoost model achieves Macro-F1 scores ranging from 0.79 to 0.94, with an average value of 0.842. These results indicate that CatBoost provides relatively strong classification performance; however, the noticeable variation across folds suggests a degree of sensitivity to data partitioning. In contrast, the standalone MLP model exhibits substantially lower and more fluctuating Macro-F1 scores, ranging from 0.33 to 0.65, with a mean value of 0.521. This wide performance variation reflects limited stability across folds and confirms that the MLP model struggles to produce consistent results when applied directly to tabular sleep disorder data. As shown in Table 2, the proposed hybrid two-stage model consistently achieves higher Macro-F1 scores across all folds, ranging from 0.87 to 0.95, with an average Macro-F1 score of 0.904. Moreover, the narrower performance range across folds indicates reduced variance and improved stability. These results demonstrate that the hybrid approach not only enhances overall classification effectiveness but also provides more robust and reliable performance compared to the standalone models. This consistency indicates that the proposed model is less affected by variations in data partitioning during cross-validation.

Table 2. Macro-F1 Scores per Fold for Each Model

Fold	CatBoost	MLP	Hybrid
1	0.805	0.410	0.909
2	0.790	0.330	0.875
3	0.945	0.650	0.873
4	0.842	0.540	0.952
5	0.839	0.402	0.901

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

3.4 Overall Performance Comparison

The overall performance comparison further confirms the effectiveness of the proposed hybrid two-stage approach. As summarized in Table 3, the standalone CatBoost model achieves an Accuracy of 0.896, a Macro-F1 score of 0.842, and an MCC of 0.791, indicating strong baseline performance on tabular sleep disorder data. These results show that CatBoost is well suited for structured medical data, although its performance is still influenced by class imbalance and data variability. In contrast, the standalone MLP model produces considerably lower results, with an Accuracy of 0.561, a Macro-F1 score of 0.521, and an MCC of 0.325, reflecting limited discriminative capability and weaker reliability in this setting. As shown in Table 3, the proposed hybrid two-stage model consistently outperforms both baselines, achieving an Accuracy of 0.941, a Macro-F1 score of 0.904, and an MCC of 0.879. Compared to the standalone CatBoost model, this represents an improvement of approximately 6.2% in Macro-F1 and 8.9% in MCC. These gains indicate not only higher overall classification accuracy but also a notable enhancement in class-balanced performance and prediction reliability. Results suggest that integrating latent feature learning with a robust tree-based classifier provides a more effective and stable solution for sleep disorder classification using tabular medical data.

Table 3. Model Result

MODEL	ACCURACY	MACRO-F1	MCC
CatBoost	0.896429	0.842179	0.790915
MLP	0.560714	0.520612	0.325198
Hybrid	0.941071	0.903632	0.879522

3.5 Breakdown of Hybrid Two-Stage Classification Results

To provide a more detailed understanding of the performance of the proposed hybrid two-stage model, a breakdown analysis is conducted using the confusion matrix, as illustrated in Figure 7. The confusion matrix in Figure 7 visualizes the distribution of correct and incorrect predictions across all classes and provides insight into the classification behavior at both the screening and subtype classification stages. At the binary screening stage, the model demonstrates excellent discrimination between healthy individuals and subjects with sleep disorders. As shown in Figure 7, all 375 healthy samples are correctly classified as Healthy, with no misclassification into insomnia or sleep apnea classes. This result confirms that the first-stage CatBoost model functions as an effective screening mechanism, successfully filtering healthy individuals before proceeding to the subtype classification stage. For the insomnia class, Figure 7 shows that the model correctly classifies 74 samples as insomnia. A limited number of insomnia cases are predicted as Healthy (11 samples) or misclassified as Sleep Apnea (8 samples). Despite these misclassifications, the majority of insomnia cases are correctly identified, indicating that the hybrid model effectively captures insomnia-related patterns. In the case of sleep apnea, the confusion matrix in Figure 7 indicates that 78 samples are correctly classified as sleep apnea. A small number of samples are misclassified as Healthy (9 samples) or Insomnia (5 samples). These results suggest that the model is able to distinguish sleep apnea cases with high reliability, while minor confusion reflects overlapping clinical characteristics between sleep disorder subtypes. Overall, the confusion matrix presented in Figure 7 confirms that the proposed hybrid two-stage framework achieves strong and balanced performance across all classes. The near-perfect classification of healthy individuals highlights the effectiveness of the screening stage, while the high correct classification rates for insomnia and sleep apnea demonstrate the reliability of the subtype classification stage. This result also indicates that the proposed hybrid two-stage model does not only perform well on the majority Healthy class, but also shows reliable performance in identifying the minority disorder classes. The remaining misclassifications between Insomnia, Sleep Apnea, and Healthy samples may occur because several sleep-related features have overlapping patterns, such as sleep duration, sleep quality, stress level, and physiological indicators. Nevertheless, the number of correct predictions for both Insomnia and Sleep Apnea remains higher than the number of misclassified samples, indicating that the model is still able to capture meaningful disorder-specific patterns. Therefore, the confusion matrix analysis supports the effectiveness of the two-stage strategy, where the first stage focuses on screening Healthy and Disorder samples, while the second stage performs more specific subtype classification. This finding further confirms that the proposed framework is effective in maintaining classification stability across both majority and minority classes.

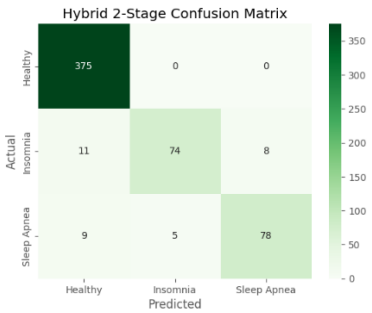

Figure 7. Confusion matrix of the hybrid two-stage model

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

3.6 Hybrid Model Robustness Evaluation

The robustness of the hybrid model is evaluated by varying the number of folds in Stratified K-Fold Cross-Validation from 5 to 25. As summarized in Table 4 and illustrated in Figure 8, the results show that the hybrid model maintains stable performance across all configurations. At 5-fold, the model achieves the highest scores with an Accuracy of 0.941, a Macro-F1 score of 0.904, and an MCC of 0.879. As the number of folds increases, performance decreases gradually but remains stable. For 10–15 folds, Accuracy ranges from 0.936 to 0.938, Macro-F1 from 0.892 to 0.896, and MCC around 0.868. For 20–25 folds, the performance stabilizes at approximately an Accuracy of 0.934, a Macro-F1 score of 0.888, and an MCC of 0.865. The absence of abrupt performance drops indicates that the hybrid model is not sensitive to a particular validation setting.

Table 4. Robustness Result

Fold	Accuracy	Macro-F1	MCC
5	0.941071	0.903632	0.879522
10	0.937500	0.896079	0.871970
15	0.935714	0.892238	0.868263
20	0.933929	0.888389	0.864567
25	0.933929	0.888389	0.864567

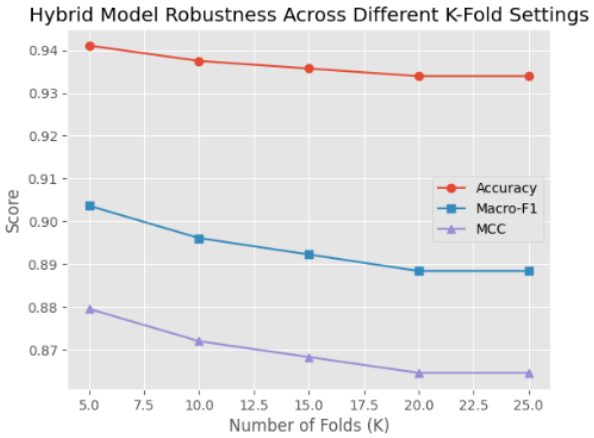

Figure 8. Robustness of the proposed hybrid model across different K-fold cross-validation settings in terms of Accuracy, Macro-F1, and MCC

3.7 Discussion

The experimental results clearly demonstrate that the proposed hybrid two-stage framework provides significant advantages over standalone models for sleep disorder classification using tabular medical data. While CatBoost alone already performs well on structured data, its capability can be further enhanced through integration with latent feature learning. Conversely, although deep learning models such as MLP are effective in capturing non-linear relationships, their performance benefits substantially from being incorporated within a structured hybrid framework rather than being applied independently. The proposed hybrid approach successfully combines these complementary strengths. The first stage functions as an effective screening mechanism that simplifies the classification task, while the second stage employs MLP-based latent feature learning to enrich feature representations. By fusing these latent features with the original input features and applying CatBoost for final classification, the proposed framework achieves an accuracy of 0.94, a Macro-F1 score of 0.90, and an MCC of 0.88, indicating strong overall performance, balanced class-wise predictions, and high prediction reliability. The consistent performance observed across multiple stratified k-fold configurations further confirms the robustness and generalization capability of the proposed approach. This stability indicates that the hybrid two-stage framework is suitable for practical sleep disorder classification scenarios involving heterogeneous and imbalanced medical tabular data.

In terms of performance comparison, the proposed Hybrid Two-Stage model achieves better results than the previous studies discussed in the introduction. The Decision Tree model reported by Ali Ridla et al. [11] achieved an accuracy of 85.33%, while the LLM-assisted machine learning pipeline proposed by Zhao [12] achieved an accuracy of 91.9%. In addition, the hybrid neural network approach developed by Airlangga [13] achieved approximately 91% accuracy, and the comparative machine learning and deep learning study by Airlangga [14] reported accuracy of up to approximately 93%. Compared with these results, the proposed Hybrid Two-Stage CatBoost and MLP model achieves the highest accuracy of 94.1%. This indicates that separating the classification process into a Healthy versus Disorder screening stage and a subtype classification stage can improve predictive performance. The results also show that combining CatBoost with MLP-based latent feature extraction provides better classification effectiveness than previous single-stage and comparative model approaches. This improvement suggests that the proposed framework is

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

not only competitive in terms of accuracy, but also more structured in handling different classification stages. Therefore, the two-stage design provides a clearer advantage by reducing classification complexity before identifying specific sleep disorder subtypes.

Overall, these comparative results indicate that the proposed framework not only improves upon traditional single-model approaches but also delivers performance that is competitive with, and in some cases superior to, more complex hybrid and deep learning-based methods. By integrating CatBoost and MLP within a two-stage design, the proposed approach offers a balanced trade-off between accuracy, stability, and practical applicability. While the present study focuses on a specific dataset and a limited number of sleep disorder subtypes, the proposed framework is inherently flexible and can be readily extended to broader datasets and additional clinical conditions. Future work may explore the incorporation of richer data sources and extended classification scopes to further enhance applicability and clinical relevance.

4. CONCLUSION

This study proposed a Hybrid Two-Stage classification framework for sleep disorder detection using medical tabular data. The framework integrates CatBoost as a binary screening model to distinguish healthy individuals from those with sleep disorders, followed by a second stage that combines MLP-based latent feature learning with CatBoost for sleep disorder subtype classification. This design enables effective handling of nonlinear feature interactions while maintaining robustness on structured data. Experimental results demonstrate that the proposed hybrid approach consistently outperforms standalone CatBoost and MLP models. The hybrid two-stage model achieves an accuracy of 94.1%, a Macro-F1 score of 0.90, and an MCC of 0.88, indicating improved overall correctness, balanced class-wise performance, and strong predictive reliability. In addition, the robustness evaluation across multiple stratified k-fold configurations confirms that the model maintains stable performance under different validation settings, highlighting its generalization capability. The results also show that addressing class imbalance using SMOTE-ENN and incorporating hyperparameter optimization into the training pipeline contribute to the overall effectiveness of the proposed framework. These components allow the model to better handle imbalanced medical data while avoiding overfitting and unstable predictions. Overall, the proposed Hybrid Two-Stage approach provides a practical and non-invasive alternative for sleep disorder classification based on tabular health data. Future work may focus on extending the framework to additional sleep disorder categories, integrating longitudinal or multimodal data, and validating the approach on larger and more diverse clinical datasets. The main contribution of this research is the proposed integration of hierarchical classification, latent feature representation, imbalance handling, and automated hyperparameter optimization in a single framework, which provides a more focused and stable approach than previous single-stage sleep disorder classification models.

REFERENCES

[1] G. Medic, M. Wille, and M. Hemels, “Short- and long-term health consequences of sleep disruption,” *Nat. Sci. Sleep*, vol. 9, pp. 151–161, May 2017, doi: 10.2147/NSS.S134864.

[2] R. Alazaidah, G. Samara, M. Aljaidi, M. Haj Qasem, A. Alsarhan, and M. Alshammari, “Potential of Machine Learning for Predicting Sleep Disorders: A Comprehensive Analysis of Regression and Classification Models,” *Diagnostics*, vol. 14, no. 1, p. 27, Dec. 2023, doi: 10.3390/diagnostics14010027.

[3] M. Mostafa Monowar *et al.*, “Advanced sleep disorder detection using multi-layered ensemble learning and advanced data balancing techniques,” *Front. Artif. Intell.*, vol. 7, Jan. 2025, doi: 10.3389/frai.2024.1506770.

[4] A. A. Alhussan, E.-S. M. El-Kenawy, D. S. Khafaga, and M. M. Eid, “Enhancing sleep disorder classification using machine learning and metaheuristic optimization techniques,” *Biomed. Signal Process. Control*, vol. 110, p. 108204, Dec. 2025, doi: 10.1016/j.bspc.2025.108204.

[5] N. Nuraeni and M. Faisal, “Classification of Sleep Disorders Using Support Vector Machine,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Jan. 2025, doi: 10.22219/kinetik.v10i1.2054.

[6] T. Misriati and R. Aryanti, “Optimasi klasifikasi gangguan tidur pada dataset tidak seimbang menggunakan SMOTE dan algoritma machine learning,” Jurnal Teknoinfo, vol. 19, no. 2, pp. 128–136, Jul. 2025, doi: 10.33365/teknoinfo.v19i2.295.

[7] T. U. Wara, A. H. Fahad, A. S. Das, and M. M. H. Shawon, “A systematic review on sleep stage classification and sleep disorder detection using artificial intelligence,” *Heliyon*, vol. 11, no. 12, p. e43576, Jul. 2025, doi: 10.1016/j.heliyon.2025.e43576.

[8] E. Dritsas and M. Trigka, “Utilizing Multi-Class Classification Methods for Automated Sleep Disorder Prediction,” *Information*, vol. 15, no. 8, p. 426, Jul. 2024, doi: 10.3390/info15080426.

[9] M. Hamid, F. Hajjej, A. S. Alluhaidan, and N. W. bin Mannie, “Fine tuned CatBoost machine learning approach for early detection of cardiovascular disease through predictive modeling,” *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-13790-x.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.

[11] M. A. Ridla and A. A. Bawazin, “Klasifikasi gangguan tidur berdasarkan gaya hidup menggunakan algoritma Decision Tree,” Journal of Information System and Application Development, vol. 3, no. 2, pp. 118–125, 2025, doi: 10.26905/jisad.v3i2.16038.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 590−600
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9958

[12] Y. Zhao, "Automatic Sleep Disorder Classification Using Large Language Model Prompting on Sleep Health and Lifestyle Data," Mar. 03, 2025. doi: 10.21203/rs.3.rs-6124845/v1.
[13] G. Airlangga, "Evaluating hybrid neural network architectures for predicting sleep disorders from structured data," JIKO (Jurnal Informatika dan Komputer), vol. 7, no. 1, pp. 58–64, Apr. 2024, doi: 10.33387/jiko.v7i1.7873.
[14] G. Airlangga, "Evaluating machine learning models for predicting sleep disorders in a lifestyle and health data context," JIKO (Jurnal Informatika dan Komputer), vol. 7, no. 1, pp. 51–57, Apr. 2024, doi: 10.33387/jiko.v7i1.7870.
[15] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, Boston, MA: Springer US, 2009, pp. 532–538. doi: 10.1007/978-0-387-39940-9_565.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 6639–6649.
[18] H.-T. Cheng et al., "Wide & deep learning for recommender systems," in Proc. 1st Workshop on Deep Learning for Recommender Systems, 2016, pp. 7–10, doi: 10.1145/2988450.2988454.
[19] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002, doi: 10.1145/505282.505283.
[20] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proc. 23rd Int. Conf. Mach. Learn. (ICML '06), New York, NY, USA: ACM Press, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.
[21] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.