Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 3, pp. 335-351, August 2026

RESEARCH PAPER

OPEN ACCESS

Depression Level Classification Using Compact Cross-Domain Feature Engineering on Sleep, Physical Activity, and Demographic Data

Nila Yoga Tama Nurwati, Fatma Indriani, Friska Abadi, Dodon Turianto Nugrahadi, and Rudy Herteno

Department of Computer Science, Faculty of Mathematics and Natural Science, Lambung Mangkurat University, Banjarbaru, Indonesia

Abstract

Depression is a common mental health disorder and a major public health concern, and early identification of depressive symptoms using population survey data can support exploratory risk analysis. However, many previous studies formulated depression prediction as a binary classification task. They used broad predictor sets, while multiclass depression-level classification with compact and interpretable cross-domain features remains less explored. This study developed a compact cross-domain feature engineering approach for classifying depression levels using sleep, physical activity, and demographic data from NHANES 2017–2018. A total of 5,068 respondents were included after preprocessing and PHQ-9 label construction. The target variable was divided into three classes: no-to-minimal depression, mild depression, and depression. Twenty raw predictors were transformed into 15 engineered features representing sleep patterns, sleep-related problems, physical activity, sedentary behavior, and interactions with age and income. Logistic Regression with class_weight = balanced was evaluated using stratified 5-fold cross-validation and compared with several baseline classifiers. The Final 15 FE Only scenario achieved an accuracy of 0.6215 ± 0.0091, macro F1-score of 0.4501 ± 0.0104, balanced accuracy of 0.5146 ± 0.0179, and depression-class recall of 0.6122 ± 0.0622. Compared with Raw Features, depression-class recall increased from 0.5360 ± 0.0541 to 0.6122 ± 0.0622, although the improvement was not statistically significant. These findings indicate that compact cross-domain features can improve sensitivity toward the depression class in an interpretable Logistic Regression setting, but overall predictive gains remain modest. The proposed model is more suitable for exploratory and population-level screening support rather than a stand-alone clinical diagnosis.

Paper History

Received April 02, 2026
Revised May 31, 2026
Accepted May 31, 2026
Published July 10, 2026

Keywords

Depression level classification;
Cross-domain feature engineering;
Logistic Regression;
PHQ-9;
Sleep patterns;
Physical activity

Author Email

nla97742@gmail.com
f.indriani@ulm.ac.id
friska.abadi@ulm.ac.id
dodonturianto@ulm.ac.id
rudy.herteno@ulm.ac.id

I. Introduction

Depression is a common mental health disorder that contributes substantially to the global burden of disease and disability [1]. This condition is not only associated with changes in mood but can also affect emotional regulation, cognitive function, social interaction, quality of life, and individual productivity [2]. In epidemiological studies, the severity of depressive symptoms is often measured using the Patient Health Questionnaire-9 (PHQ-9), as this instrument has good reliability and validity in identifying depressive symptoms across various populations [3][4]. One lifestyle factor that has been widely associated with depression is sleep patterns. Sleep plays an important role in maintaining physical and psychological balance. Several studies have shown that non-optimal sleep duration, either too short or too long, is associated with an increased risk of depression [5]. In addition, poor sleep quality, sleep disorders such as insomnia, and daytime sleepiness are also associated with increased depressive symptoms and poorer psychological conditions [6][7]. Population-based studies have also shown that sleep duration, sleep disorders, and sleep quality are related to depressive symptoms [8][9]. These findings indicate that sleep patterns should not be examined using only a single indicator; therefore, sleep-related information needs to be represented more appropriately in the analysis process.

In addition to sleep patterns, physical activity and sedentary behavior are also related to mental health conditions. Several studies have shown that low physical activity and high sedentary time are associated with increased depressive symptoms, whereas higher levels of physical activity are associated with a lower likelihood of depressive symptoms [10][11][12]. Similar findings have also been reported in NHANES-based studies, which showed that sedentary behavior and physical activity are associated with depressive symptoms in the adult population [13][14]. Therefore, these two factors are relevant to be considered together with sleep patterns in the analysis of depression levels.

The development of machine learning has encouraged the application of data-driven classification methods in the healthcare field, including the classification or prediction of depressive symptoms [15][16]. Several previous studies have used algorithms such as Logistic

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

Regression, Random Forest, Gradient Boosting, and other machine learning models to classify or predict depression severity based on clinical, biomarker, survey, and lifestyle-related factors [17]. In the context of national health survey data, Vu et al. [18] compared several machine learning models for depression prediction using NHANES data and reported reasonable predictive performance across the tested algorithms.

Previous NHANES-based depression prediction studies have mainly focused on binary depression classification, determinant analysis, biomarker-based prediction, explainable machine learning, or population-specific cohorts [18][19][20][21][22]. Many of these studies relied on broad sets of raw demographic, clinical, behavioral, or lifestyle variables and often emphasized predictive optimization or explainability using multiple machine learning models. However, the explicit construction of compact and interpretable interaction-based features derived from multiple domains has received comparatively limited attention, particularly in multiclass depression level classification settings using NHANES survey data.

In structured health survey data, raw variables may not optimally represent relationships among sleep patterns, physical activity, and socioeconomic conditions because these factors may involve cross-domain interactions rather than fully independent contributions. This limitation becomes more relevant in interpretable linear models such as Logistic Regression, where predictors are primarily modeled additively unless interaction representations are explicitly constructed. Furthermore, multiclass depression classification based on PHQ-9 severity levels remains challenging because adjacent categories, especially mild depression, often exhibit overlapping symptom characteristics and imbalanced class distributions [23]. Therefore, a compact cross-domain feature engineering approach may provide a more interpretable representation for examining depression-level classification in imbalanced NHANES survey data, while allowing the effect of interaction-based features on minority-class sensitivity to be evaluated in an interpretable Logistic Regression setting.

Based on these issues, this study aims to classify depression levels using a compact cross-domain feature engineering approach based on NHANES 2017–2018 data [24]. Feature engineering was performed by constructing 15 final features derived from sleep patterns, physical activity, and demographic interaction domains. Logistic Regression was used as the main interpretable baseline model, while Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost were included as baseline model comparisons. The evaluation was conducted using Stratified 5-Fold Cross Validation with accuracy, precision, recall, macro F1-score, balanced accuracy, recall and F1-score for the depression class, AUC-ROC, and PR-AUC for the depression class as evaluation metrics.

The main contributions of this study are as follows: 1) a set of 15 interpretable cross-domain features constructed from sleep pattern, physical activity, and demographic variables, reducing the input dimensionality from 20 raw features to 15 while maintaining similar classification performance; 2) an evaluation showing that the engineered features achieved higher depression-class recall in the main Logistic Regression setting, although the overall improvement remained modest and was not statistically significant; 3) a per-class and confusion-matrix analysis that characterizes where the model succeeds and fails across the three depression levels, rather than relying only on overall accuracy; and 4) an assessment of the precision-recall trade-off on imbalanced PHQ-9 data using Logistic Regression as an interpretable baseline.

This study is structured as follows. Section 2 describes the dataset, data collection, data processing, and statistical analysis. Section 3 presents the classification results, per-class performance, confusion matrix analysis, and baseline model comparison. Section 4 discusses the interpretation of the results, the impact of cross-domain feature engineering, limitations, clinical implications, and future research directions. Section 5 presents the conclusion of this study.

II. Materials and Method

A. Dataset and Data Collection

This study used the National Health and Nutrition Examination Survey (NHANES) 2017–2018 dataset provided by the Centers for Disease Control and Prevention (CDC) [24]. NHANES is a national-scale health survey that collects demographic, health, lifestyle, physical activity, sleep, and mental health information. The dataset used in this study was obtained from publicly available secondary data; therefore, no primary data were collected directly from respondents. Data were obtained from four main components: Demographics Data (DEMO), Sleep Disorders Questionnaire (SLQ), Physical Activity Questionnaire (PAQ), and Depression Questionnaire (DPQ). The DEMO component was used to obtain demographic information, including age, sex, race/ethnicity, education level, marital status, family income-to-poverty ratio, household size, and country of birth. The SLQ component provided sleep-related information, while the PAQ component provided physical activity and sedentary behavior information. The DPQ component provided the nine PHQ-9 items used to construct the depression level labels.

The initial data consisted of 9,254 records in DEMO, 6,161 records in SLQ, 5,856 records in PAQ, and 5,533 records in DPQ. All data components were merged using the unique respondent identifier, SEQN. After merging the relevant variables, 5,533 records with 40 variables were obtained. The target variable was constructed from the total Patient Health Questionnaire-9 (PHQ-9) score derived from the nine DPQ questionnaire items [3][4]. Respondents with missing values or invalid responses in the PHQ-9 items were removed because these items were directly used to construct the depression level labels. After this stage, the final dataset consisted of 5,068 records. Depression levels were classified into three classes based on the total PHQ-9 score. The no-to-minimal depression class represented a score of 0–4, the

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

mild depression class represented a score of 5–9, and the depression class represented a score of ≥10 [25][26]. The final class distribution consisted of 3,772 respondents in the no-to-minimal depression class, or 74.4%; 837 respondents in the mild depression class, or 16.5%; and 459 respondents in the depression class, or 9.1%. This distribution indicates that the dataset used in this study was imbalanced.

B. Data Processing

Data processing in this study consisted of two main stages: initial data cleaning and cross-domain feature engineering. Initial data cleaning was performed to prepare the NHANES data and construct the target label, whereas cross-domain feature engineering was used to derive compact and interpretable features from the sleep, physical activity, and demographic domains. These stages are important in data modeling because data preprocessing, feature engineering, missing-value handling, and model evaluation can affect data quality, model performance, reproducibility, and interpretability [27][28][29][30]. The NHANES 2017–2018 dataset was constructed by merging four modules, namely demographic data (DEMO), sleep questionnaire data (SLQ), physical activity questionnaire data (PAQ), and depression questionnaire data (DPQ), using the respondent identifier SEQN. The PHQ-9 total score was calculated from the nine DPQ items and used to construct three depression level labels: no-to-minimal depression, mild depression, and depression. Respondents with missing or invalid PHQ-9 item responses were excluded because the DPQ items were directly used to define the target label. The PHQ-9 total score for respondent i was calculated as shown in Eq. (1):

$$PHQ9_i = \sum_{j=1}^{9} DPQ_{ij} \quad (1)$$

where $PHQ9_i$ denotes the total PHQ-9 score for respondent i, and DPQ_{ij} denotes the score of the j-th PHQ-9 item for respondent i. Based on the total PHQ-9 score, the depression level label was constructed using the threshold rule in Eq. (2):

$$\begin{aligned} y_i &= 0, if\ PHQ9_i \leq 4 \\ y_i &= 1, if\ 5 \leq PHQ9_i \leq 9 \\ y_i &= 2, if\ PHQ9_i \geq 10 \end{aligned} \quad (2)$$

where $y_i = 0$ represents no-to-minimal depression, $y_i = 1$ represents mild depression, and $y_i = 2$ represents depression [25][26].

Special NHANES response codes were handled according to the coding structure of each variable. Response codes indicating refused, don't know, or missing responses, such as 7, 9, 77, 99, 7777, and 9999, were converted into missing values only for variables in which those codes represented invalid responses. This variable-specific handling was applied to avoid incorrectly treating valid values, such as seven or nine hours of sleep, as missing data. Binary variables in the physical activity and sleep components were recoded as 1 for "Yes" responses and 0 for "No" responses. In addition, skip-logic in the physical activity variables was considered by assigning an activity duration value of 0 when respondents reported not performing the corresponding activity. Missing value handling is necessary because incomplete data can affect the quality and validity of the analysis, and the selection of an imputation strategy should be adjusted to the characteristics of the dataset [28][29]. To avoid data leakage, missing value imputation was not fitted before cross-validation. Instead, imputation was performed inside the cross-validation pipeline. Numerical variables were imputed using the median, whereas categorical variables were imputed using the mode. The imputation parameters were estimated only from the training fold and then applied to the corresponding validation fold. Numerical imputation was formulated as shown in Eq. (3):

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, if\ x_{ij}\ is\ not\ missing \\ median(X_j^{train}), if\ x_{ij}\ is\ missing \end{cases} \quad (3)$$

where $\tilde{x}_{ij}$ denotes the imputed value of feature j for respondent i, and median(X_j^{train}) denotes the median value of feature j estimated from the training fold. Categorical imputation was formulated as shown in Eq. (4):

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, if\ x_{ij}\ is\ not\ missing \\ mode(X_j^{train}), if\ x_{ij}\ is\ missing \end{cases} \quad (4)$$

where mode (X_j^{train}) denotes the most frequent category of feature j estimated from the training fold.

Categorical variables in the raw feature scenarios were transformed using one-hot encoding. After feature transformation, the resulting feature matrix was standardized using z-score standardization. Similar to imputation and encoding, standardization parameters were estimated only from the training fold and then applied to the corresponding validation fold to prevent information from the validation fold from influencing the model training process. Z-score standardization was calculated using Eq. (5):

$$z_{ij} = \frac{\left(\tilde{x}_{ij} - \mu_j^{train}\right)}{\sigma_j^{train}} \quad (5)$$

where z_{ij} denotes the standardized value of feature j for respondent i, $\tilde{x}_{ij}$ denotes the imputed value, while μ_j^{train} and σ_j^{train} denote the mean and standard deviation of feature j estimated from the training fold. After the initial data cleaning and target construction, the analysis dataset consisted of 5,068 records with 20 raw predictor features. These raw features were derived from the DEMO, SLQ, and PAQ components, whereas the PHQ-9 items from the DPQ component were used only to construct the target label. Missing values in the predictor variables were handled within the cross-validation pipeline. A summary of the raw predictor features, data types, and value ranges is shown in Table 1. Based on the raw predictor features in Table 1, this study constructed 15 final engineered features. Feature engineering was performed to construct more representative features by considering the data characteristics and the relationships among variables

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

from the sleep, physical activity, and demographic domains [27]. The feature engineering process was implemented inside the cross-validation pipeline so that engineered features in each validation fold were generated using transformations fitted only from the corresponding training fold. The engineered features were grouped into sleep-related features, physical activity-related features, and demographic interaction features.

Table 1. Raw predictor features after initial data cleaning.

Raw features	Variable description (range/coding)	Variable type
RIDAGEYR	Age (18–80 years)	Numeric
RIAGENDR	Gender (1–2)	Binary categorical
RIDRETH3	Race/ethnicity (1, 2, 3, 4, 6)	Categorical
DMDEDUC2	Education level (1–5)	Ordinal categorical
DMDMARTL	Marital status (1–6)	Categorical
INDFMPIR	Income-to-poverty ratio (0.00–5.00)	Numeric
DMDHHSIZ	Household size (1–6)	Numeric
DMDBORN4	Country of birth (1, 2)	Categorical
SLD012	Weekday sleep duration (2–14 hours)	Numeric
SLD013	Weekend sleep duration (2–14 hours)	Numeric
SLQ030	Snoring frequency (0–3)	Ordinal categorical
SLQ040	Breathing interruption frequency (0–3)	Ordinal categorical
SLQ050	Sleep problem indicator (0 = No, 1 = Yes)	Binary
SLQ120	Daytime sleepiness frequency (0–4)	Ordinal categorical
PAQ605	Vigorous work activity (0 = No, 1 = Yes)	Binary
PAQ620	Moderate work activity (0 = No, 1 = Yes)	Binary
PAQ635	Walking/bicycling for transport (0 = No, 1 = Yes)	Binary
PAQ650	Vigorous recreational activity (0 = No, 1 = Yes)	Binary
PAQ665	Moderate recreational activity (0 = No, 1 = Yes)	Binary
PAD680	Sedentary time (0–1320 minutes/day)	Numeric

In the sleep domain, sleep_avg was constructed from the average sleep duration on weekdays (SLD012) and weekends (SLD013), whereas sleep_diff represented the absolute difference between weekend and weekday sleep duration. These features were used to represent sleep duration patterns and weekday-weekend sleep differences, which have been associated with depressive symptoms in previous studies [5][31]. The sleep_avg and sleep_diff features were formulated in Eq. (6) and Eq. (7).

$$sleep\_avg_i = \frac{SLD012_i + SLD013_i}{2} \quad (6)$$

$$sleep\_diff_i = |SLD013_i - SLD012_i| \quad (7)$$

The sleep_flag feature was used as an indicator of non-optimal sleep duration. This feature was assigned a value of 1 if sleep_avg$_i$ < 7 or sleep_avg$_i$ > 9, and a value of 0 otherwise, as shown in Eq. (8).

$$\text{sleep\_flag}_i = \begin{cases} 1, \text{if sleep\_avg}_i < 7 \text{ or sleep\_avg}_i > 9 \\ 0, \text{if } 7 \le \text{sleep\_avg}_i \le 9 \end{cases} \quad (8)$$

The sleep_quality_score feature was constructed by summing several sleep problem indicators, namely SLQ030, SLQ040, SLQ050, and SLQ120, as formulated in Eq. (9):

$$SQS_i = SLQ030_i + SLQ040_i + SLQ050_i + SLQ120_i \quad (9)$$

where SQS_i denotes sleep_quality_score.

The sleep_problem_flag feature was constructed directly from SLQ050 as an indicator of sleep problems, as shown in Eq. (10):

$$\text{sleep\_problem\_flag}_i = SLQ050_i \quad (10)$$

These sleep-related features were included because non-optimal sleep duration and sleep-related problems have been associated with depressive symptoms in previous studies [6][8][31].

To represent the interaction between sleep-related symptoms, socioeconomic conditions, and social isolation status, this study constructed the sleepiness_poverty and sleep_isolation features. The sleepiness_poverty feature combines daytime sleepiness with the family income-to-poverty ratio, considering that sleep-related problems and income-related disadvantage have been associated with depressive symptoms [31][32]. This feature was formulated in Eq. (11):

$$\text{sleepiness\_poverty}_i = \frac{SLQ120_i}{INDFMPIR_i + 0.5} \quad (11)$$

The constant 0.5 in Eq. (11) was used to avoid division by zero when the income-to-poverty ratio was very low. The sleep_isolation feature combines sleep-related problem severity with social isolation status, considering that social isolation has been associated with depressive symptoms and sleep-related problems [31][33]. This feature was formulated in Eq. (12):

$$\text{sleep\_isolation}_i = \text{sleep\_quality\_score}_i \times \text{is\_alone}_i \quad (12)$$

where is_alone$_i$ was assigned a value of 1 if the respondent was widowed, divorced, separated, or never married, and a value of 0 otherwise.

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

In the physical activity domain, activity_count was constructed from the sum of five physical activity indicators, namely PAQ605, PAQ620, PAQ635, PAQ650, and PAQ665. This feature was used to summarize the respondent's participation in physical activity, as physical activity has been associated with depressive symptoms in an NHANES-based study [14]. The activity_count feature was formulated in Eq. (13).

$$\text{activity\_count}_i = \text{PAQ605}_i + \text{PAQ620}_i + \text{PAQ635}_i + \text{PAQ650}_i + \text{PAQ665}_i \quad (13)$$

The active_flag feature was assigned a value of 1 if activity_count$_i$ > 0, and a value of 0 otherwise, as formulated in Eq. (14):

$$active\_flag_i = \begin{cases} 1, if\ activity\_count_i > 0 \\ 0, if\ activity\_count_i = 0 \end{cases} \quad (14)$$

Sedentary behavior was represented through sedentary_ratio and sedentary_flag. The sedentary_ratio feature was constructed from the ratio of sedentary duration (PAD680) to average sleep duration converted into minutes. This feature was included because sedentary behavior has been associated with depressive symptoms and may provide complementary information to physical activity indicators in an NHANES-based study [14]. A recent systematic review and meta-analysis also reported that total sedentary behavior was associated with a higher risk of depression, supporting the relevance of including sedentary-related representations in this study [34]. The sedentary_ratio feature is formulated in Eq. (15).

$$\text{sedentary\_ratio}_i = \frac{\text{PAD680}_i}{(\text{sleep\_avg}_i \times 60) + 1} \quad (15)$$

The constant 1 in Eq. (15) was used to avoid division by zero. The sedentary_flag feature was assigned a value of 1 if PAD680$_i$ > 600, and a value of 0 otherwise, as shown in Eq. (16):

$$sedentary\_flag_i = \begin{cases} 1, \text{if PAD680}_i > 600 \\ 0, \text{if PAD680}_i \leq 600 \end{cases} \quad (16)$$

These features were used to summarize sedentary behavior relative to sleep duration, rather than treating sedentary time only as a standalone raw variable.

In the demographic interaction domain, the engineered features were constructed by combining age (RIDAGEYR) and family income-to-poverty ratio (INDFMPIR) with sleep and physical activity features. These interaction features were used to capture whether sleep and activity patterns may provide different information across age and socioeconomic conditions. Age, income-related factors, sleep characteristics, and physical activity have been associated with depressive symptoms in previous NHANES-based studies [14][31][32]. The age_sleep_interact and age_activity_interact features were constructed from the interaction between age and average sleep duration, and between age and the number of physical activities, respectively. To keep the equations concise, these interaction features were abbreviated as ASI and AAI. These features were formulated in Eq. (17) and Eq. (18).

$$ASI_i = \text{RIDAGEYR}_i \times \text{sleep\_avg}_i \quad (17)$$

$$\text{AAI}_i = \text{RIDAGEYR}_i \times \text{activity\_count}_i \quad (18)$$

Furthermore, income_activity_interact and income_sleep_interact were constructed to represent the interaction between socioeconomic conditions and physical activity and sleep patterns. These interaction features were abbreviated as IAI and ISI, as formulated in Eq. (19) and Eq. (20).

$$\text{IAI}_i = \text{INDFMPIR}_i \times \text{activity\_count}_i \quad (19)$$

$$\text{ISI}_i = \text{INDFMPIR}_i \times \text{sleep\_avg}_i \quad (20)$$

where ASI_i, AAI_i, IAI_i, and ISI_i denote age_sleep_interact, age_activity_interact, income_activity_interact, and income_sleep_interact for respondent i, respectively.

After all engineered features were constructed, the 15 final engineered features used in the model evaluation were summarized based on their feature groups, as shown in Table 2.

Table 2. Final engineered features used in this study.

Feature group	Engineered features
Sleep-related features	sleep_avg, sleep_diff, sleep_flag, sleep_quality_score, sleep_problem_flag, sleepiness_poverty, sleep_isolation
Activity-related features	activity_count, active_flag, sedentary_ratio, sedentary_flag
Demographic interaction features	age_sleep_interact, age_activity_interact, income_activity_interact, income_sleep_interact

The overall research workflow is illustrated in Fig. 1. To improve transparency and reproducibility in reporting the prediction pipeline, the complete experimental procedure used in this study is summarized in Algorithm 1 [35].

Algorithm 1. Depression level classification pipeline using compact cross-domain feature engineering.

Input: NHANES 2017–2018 DEMO, SLQ, PAQ, and DPQ datasets

Output: Depression level classification results and evaluation metrics

1. Load and merge DEMO, SLQ, PAQ, and DPQ datasets using SEQN.
2. Select demographic, sleep, physical activity, and PHQ-9 variables.
3. Calculate the PHQ-9 total score and construct the three-class depression label.
4. Remove respondents with missing or invalid PHQ-9 item responses.
5. Define three feature scenarios: Raw Features, Raw + Final 15 FE, and Final 15 FE Only.
6. Apply Stratified 5-Fold Cross Validation.
7. For each fold:
   a. Fit imputation, encoding, standardization, and imbalance-handling procedures using only the

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

training fold.

b. Construct the engineered features within the fold.
c. Apply the fitted transformations to the validation fold.
d. Train the classifier using the training fold.
e. Predict depression labels and class probabilities on the validation fold.

8. Calculate evaluation metrics for each fold.
9. Average the evaluation results across the five folds.
10. Perform Wilcoxon signed-rank testing to compare feature scenarios.
11. Report the final classification performance, per-class results, confusion matrix, baseline model comparison, and imbalance-handling sensitivity analysis.

C. Statistical Analysis

The main classification model used in this study was multinomial Logistic Regression. Logistic Regression was selected because it is suitable for tabular health survey data, can be applied to multiclass classification, and provides an interpretable baseline model. Recent methodological discussions also suggest that the performance advantage of more complex machine learning models over Logistic Regression in structured clinical tabular data is often context-dependent and influenced by dataset characteristics and data quality [36]. Interpretability is important in healthcare-related machine learning because model outputs need to be understandable and clinically reviewable [37]. In this study, Logistic Regression was used to classify three target classes: no-to-minimal depression, mild depression, and depression. For respondent i with feature vector x_i, the multinomial Logistic Regression model estimates the probability of class k using the softmax function, as shown in Eq. (21):

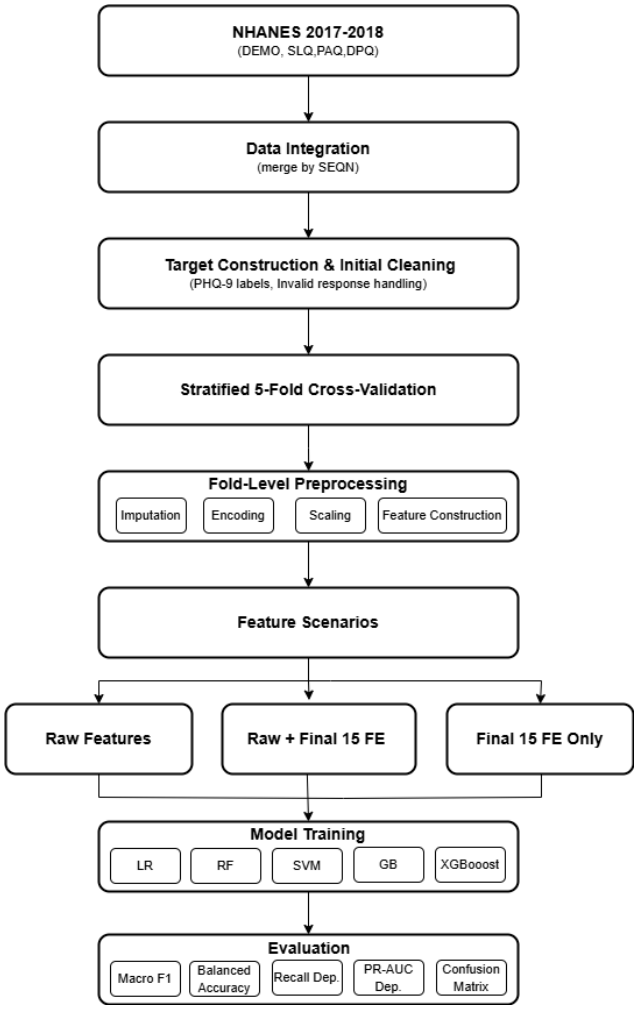

Fig. 1. Study workflow for depression level classification using cross-domain feature engineering.

$$p_{ik} = \frac{exp(w_k^T xi + b_k)}{\sum_{c=0}^{2} exp(w_c^T x_i + b_c)} \quad (21)$$

where p_{ik} denotes the predicted probability that respondent i belongs to class k, x_i denotes the feature vector of respondent i, w_k denotes the coefficient vector for class k, b_k denotes the intercept for class k, and $k \in \{0,1,2\}$ represents the three depression level classes. The predicted class was determined using the maximum predicted probability, as formulated in Eq. (22):

$$\hat{y}_i = \arg\max_k p_{ik}, k \in \{0,1,2\} \quad (22)$$

where $\hat{y}_i$ denotes the predicted class label for respondent i.

The Logistic Regression model was implemented using the multinomial setting with the lbfgs solver, L2 regularization, C = 1.0, max_iter = 2000, and class_weight = balanced. The class-weighting strategy was used because the dataset had an imbalanced class distribution, with the depression class representing the smallest group. This setting assigns higher weights to minority classes during training, reducing the tendency of the model to be dominated by the majority class [38]. The balanced class weight for class k was calculated using Eq. (23):

$$cw_k = \frac{N}{K \times n_k} \quad (23)$$

where cw_k denotes the class weight for class k, N denotes the total number of training samples, K denotes the number of classes, and n_k denotes the number of training samples in class k. The training objective of Logistic Regression was based on weighted cross-entropy loss. The loss function used to optimize the model parameters can be expressed as Eq. (24):

$$L = \frac{-1}{N} \sum_{i=1}^{N} [\sum_{k=0}^{2} cw_k I(y_i = k) \log(p_{ik})] + \lambda ||W||_2^2 \quad (24)$$

where L denotes the regularized weighted loss, N denotes the number of training samples, cw_k denotes the class weight for class k, $I(y_i = k)$ is an indicator function that equals 1 when the true class of respondent i is k and 0 otherwise, p_{ik} denotes the predicted probability for class k, λ denotes the regularization parameter, and $||W||_2^2$ denotes the L2 penalty term.

Model training and evaluation were conducted using three feature scenarios: Raw Features, Raw + Final 15 FE, and Final 15 FE Only. The Raw Features scenario

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

used the 20 original predictor variables from the demographic, sleep, and physical activity domains. The Final 15 FE Only scenario used only the proposed engineered features, while the Raw + Final 15 FE scenario combined the raw and engineered features. This comparison was conducted to examine whether the proposed engineered features provided a compact and informative representation compared with raw predictors and combined features. To strengthen the evaluation, Logistic Regression was also compared with several baseline machine learning models, namely Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost. All models were evaluated using the same feature scenarios and cross-validation setting. Logistic Regression, Random Forest, and Support Vector Machine were applied with class balancing, whereas Gradient Boosting and XGBoost used balanced sample weights during training. This comparison was used as a baseline evaluation, while extensive hyperparameter optimization was not performed because the main focus of this study was compact cross-domain feature engineering with interpretable modeling. Model evaluation was conducted using Stratified 5-Fold Cross Validation. Stratification was used to maintain the proportion of each depression class in every fold, considering the imbalanced class distribution in the dataset [39]. In each fold, all preprocessing steps, including imputation, one-hot encoding, feature construction, standardization, imbalance handling, and model fitting, were performed using only the training fold. The fitted preprocessing transformations were then applied to the corresponding validation fold. This procedure was applied to avoid data leakage and reduce overly optimistic performance estimates [40]. For each evaluation metric, the average performance across the five folds was calculated using Eq. (25):

$$M_{mean} = \frac{\sum_{f=1}^{5} M_f}{5} \tag{25}$$

where M_{mean} denotes the average metric value across folds, and M_f denotes the metric value obtained from fold f. The standard deviation across folds was calculated using Eq. (26):

$$M_{SD} = \sqrt{\frac{\sum_{f=1}^{5}(M_f - M_{mean})^2}{5}} \tag{26}$$

where M_{SD} denotes the standard deviation of the metric values across the five folds. The evaluation metrics included accuracy, macro precision, macro recall, macro F1-score, balanced accuracy, AUC-ROC, precision for the depression class, recall for the depression class, F1-score for the depression class, and PR-AUC for the depression class. Macro F1-score was used as the main overall metric because it assigns equal weight to each class and is more appropriate than accuracy alone for imbalanced multiclass classification. Balanced accuracy was used to evaluate the average recall across classes. AUC-ROC was calculated using a multiclass one-vs-rest macro-average approach, while PR-AUC was reported for the depression class because precision-recall-based evaluation is useful for assessing minority-class performance [38][30]. Accuracy was calculated as shown in Eq. (27):

$$Accuracy = \frac{\sum_{k=0}^{2} TP_k}{N} \tag{27}$$

where TP_k denotes the number of correctly classified samples in class k, and N denotes the total number of samples. For each class k, precision, recall, and F1-score were calculated using Eq. (28), Eq. (29), and Eq. (30), respectively:

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{28}$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{29}$$

$$F_{1k} = \frac{2 \times Precision_k \times Recall_k}{Precision_k + Recall_k} \tag{30}$$

where TP_k, FP_k, and FN_k denote true positives, false positives, and false negatives for class k, respectively. Macro F1-score was calculated by averaging the F1-score of all classes, as shown in Eq. (31):

$$Macro\ F1 = \frac{1}{K}\sum_{k=0}^{2} F1_k \tag{31}$$

Balanced accuracy was calculated by averaging the recall values across all classes, as formulated in Eq. (32):

$$Balanced\ Accuracy = \frac{1}{K}\sum_{k=0}^{2} Recall_k \tag{32}$$

where K denotes the number of classes. In this study, K = 3. To provide an additional paired comparison between feature scenarios, this study applied the Wilcoxon signed-rank test to fold-level results [30]. The Wilcoxon signed-rank test was used to compare the Final 15 FE Only scenario against the Raw Features scenario and the Raw + Final 15 FE scenario for Logistic Regression. The tested metrics included macro F1-score, balanced accuracy, recall for the depression class, F1-score for the depression class, and PR-AUC for the depression class. For each fold, the difference between paired metric values was calculated using Eq. (33):

$$d_f = M_{FE,f} - M_{baseline,f} \tag{33}$$

where d_f denotes the fold-level difference between the Final 15 FE Only scenario and the compared baseline scenario for fold f. The Wilcoxon signed-rank statistic was then computed from the ranks of the absolute non-zero differences, as shown in Eq. (34):

$$W = min(W^+, W^-) \tag{34}$$

where W^+ denotes the sum of ranks for positive differences and W^- denotes the sum of ranks for negative differences. A p-value below 0.05 was considered statistically significant. Because class imbalance was a central issue in the dataset, an additional sensitivity analysis was conducted using Logistic Regression with the Final 15 FE Only scenario. Four imbalance-handling strategies were compared: no imbalance handling,

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

class_weight = balanced, SMOTE, and random undersampling [38][41]. SMOTE and random undersampling were applied only to the training fold after imputation, feature construction, and standardization. The validation fold was not resampled. The Wilcoxon signed-rank test was also applied to compare class_weight = balanced with no imbalance handling, SMOTE, and random undersampling based on fold-level results. This analysis was conducted to examine whether the selected class-weighting strategy remained reasonable compared with other common imbalance-handling methods, without shifting the main focus of the study toward imbalance-handling optimization.

III. Results

A. Model Performance

Depression level classification testing was first conducted using Logistic Regression with the Final 15 FE Only scenario. The evaluation was performed using Stratified 5-Fold Cross Validation to maintain the proportion of each depression category in the training and validation folds. The main performance of Logistic Regression using the 15 final engineered features is shown in Table 3.

Table 3. Performance of Logistic Regression using 15 final engineered features.

Metric	Mean ± SD
Accuracy	0.6215 ± 0.0091
Macro Precision	0.4459 ± 0.0112
Macro Recall	0.5146 ± 0.0179
Macro F1-score	0.4501 ± 0.0104
Balanced Accuracy	0.5146 ± 0.0179
AUC-ROC	0.7271 ± 0.0155
Precision for Depression Class	0.2390 ± 0.0162
Recall for Depression Class	0.6122 ± 0.0622
F1-score for Depression Class	0.3436 ± 0.0256
PR-AUC for Depression Class	0.2665 ± 0.0393

Based on Table 3, Logistic Regression with the Final 15 FE Only scenario achieved an accuracy of 0.6215 ± 0.0091, a macro F1-score of 0.4501 ± 0.0104, and a balanced accuracy of 0.5146 ± 0.0179. These results indicate that the overall performance remained limited in the three-class classification setting. However, the recall for the depression class reached 0.6122 ± 0.0622, showing that the model was able to identify a considerable proportion of respondents in the depression category. In contrast, the precision for the depression class was only 0.2390 ± 0.0162, indicating that many respondents predicted as depression were actually from other classes. Therefore, the model demonstrated relatively higher sensitivity toward the depression class, but its precision remained limited. To examine the effect of feature engineering, Logistic Regression performance was compared across three feature scenarios: Raw Features, Raw + Final 15 FE, and Final 15 FE Only. The performance comparison of the three scenarios is shown in Table 4.

Table 4. Comparison of raw features and final engineered features using Logistic Regression.

Metric	Raw Features	Raw + Final 15 FE	Final 15 FE Only
Features	20	35	15
Accuracy	0.6064 ± 0.0148	0.6048 ± 0.0128	**0.6215 ± 0.0091**
Macro Precision	0.4380 ± 0.0144	0.4372 ± 0.0113	**0.4459 ± 0.0112**
Macro Recall	0.4929 ± 0.0231	0.4903 ± 0.0202	**0.5146 ± 0.0179**
Macro F1-score	0.4400 ± 0.0151	0.4383 ± 0.0104	**0.4501 ± 0.0104**
Balanced Accuracy	0.4929 ± 0.0231	0.4903 ± 0.0202	**0.5146 ± 0.0179**
AUC-ROC	**0.7289 ± 0.0091**	0.7261 ± 0.0081	0.7271 ± 0.0155
Precision for Depression Class	0.2267 ± 0.0066	0.2198 ± 0.0137	**0.2390 ± 0.0162**
Recall for Depression Class	0.5360 ± 0.0541	0.5230 ± 0.0756	**0.6122 ± 0.0622**
F1-score for Depression Class	0.3183 ± 0.0160	0.3089 ± 0.0267	**0.3436 ± 0.0256**
PR-AUC for Depression Class	**0.2756 ± 0.0246**	0.2602 ± 0.0270	0.2665 ± 0.0393

Based on Table 4, the Final 15 FE Only scenario achieved the highest macro F1-score among the Logistic Regression feature scenarios, with a value of 0.4501 ± 0.0104. This value was slightly higher than Raw Features at 0.4400 ± 0.0151 and Raw + Final 15 FE at 0.4383 ± 0.0104. The Final 15 FE Only scenario also achieved the highest balanced accuracy, recall for the depression class, and F1-score for the depression class. Compared with Raw Features, the recall for the depression class increased from 0.5360 ± 0.0541 to 0.6122 ± 0.0622, while the F1-score for the depression class increased from 0.3183 ± 0.0160 to 0.3436 ± 0.0256. These results suggest that the engineered features mainly improved sensitivity toward the depression class in the Logistic Regression setting.

However, the improvement should be interpreted carefully. The precision for the depression class remained low at 0.2390 ± 0.0162, indicating that the gain was mainly reflected in recall rather than in a balanced precision-recall improvement. In addition, the PR-AUC for the depression class did not improve compared with Raw Features, which achieved 0.2756 ± 0.0246, while the Final

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

15 FE Only scenario achieved 0.2665 ± 0.0393. Therefore, the engineered features provided a compact representation and improved depression-class sensitivity, but they did not produce a large overall predictive improvement. To further examine depression-class performance, the comparison of recall and F1-score for the depression class across the three feature scenarios is shown in Fig. 2.

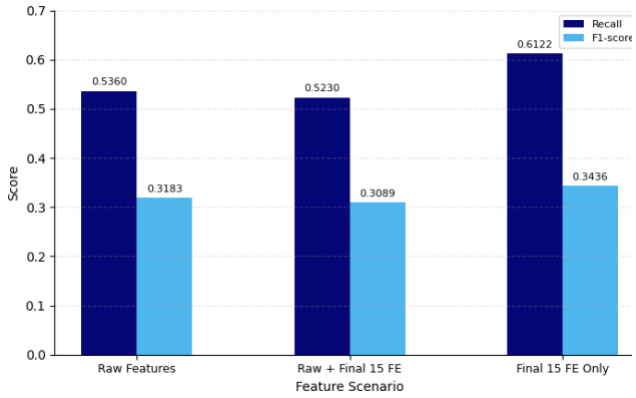

Fig. 2. Recall and F1-score for the depression

As shown in Fig. 2, the Final 15 FE Only scenario produced the highest depression-class recall and F1-score among the Logistic Regression feature scenarios. This result supports the interpretation that the proposed engineered features mainly improved sensitivity toward the depression class, although the overall gain remained modest. Fold-level statistical testing using the Wilcoxon signed-rank test showed that the improvement of Final 15 FE Only over Raw Features was not statistically significant at the 0.05 level. The p-value was 0.3125 for macro F1-score, 0.0625 for recall for the depression class, and 0.0625 for F1-score for the depression class. Therefore, the performance gain should be interpreted as a modest improvement rather than a statistically significant increase. Overall, the Final 15 FE Only scenario was retained as the main Logistic Regression result because it provided a compact 15-feature representation and achieved the highest recall and F1-score for the depression class among the Logistic Regression feature scenarios, although the overall predictive performance remained limited.

B. Per-Class Performance

Per-class performance analysis was conducted to examine the model's ability to identify each depression level. The F1-score results for each class using Logistic Regression with the Final 15 FE Only scenario are shown in Table 5.

Table 5. F1-score of each depression level using Logistic Regression with Final 15 FE.

Depression level	F1-score
No-to-minimal depression	0.7819 ± 0.0105
Mild depression	0.2250 ± 0.0345
Depression	0.3436 ± 0.0256

Based on Table 5, Logistic Regression with the Final 15 FE Only scenario achieved the highest F1-score for the no-to-minimal depression class, with a value of 0.7819 ± 0.0105. This result indicates that the model was more capable of recognizing the majority class, which is consistent with the class distribution in the dataset. In contrast, the mild depression class obtained the lowest F1-score, with a value of 0.2250 ± 0.0345. This shows that mild depression was the most difficult class to identify. The low performance of the mild depression class may be related to its position as an intermediate category between no-to-minimal depression and depression. Respondents in this class may share overlapping symptom patterns with both neighboring classes, especially when only sleep, physical activity, and demographic predictors are used. Therefore, the available predictors may not be sufficiently discriminative to separate mild depression from the other depression levels clearly. For the depression class, the model achieved an F1-score of 0.3436 ± 0.0256. Although this value remained limited, the recall for the depression class reached 0.6122 ± 0.0622, as reported in Table 3. This indicates that the model was relatively sensitive in identifying respondents in the depression category, but this sensitivity was not accompanied by high precision. Overall, the per-class results show that the model performed best on the majority class, struggled most with mild depression, and showed moderate sensitivity toward the depression class.

C. Confusion Matrix Analysis

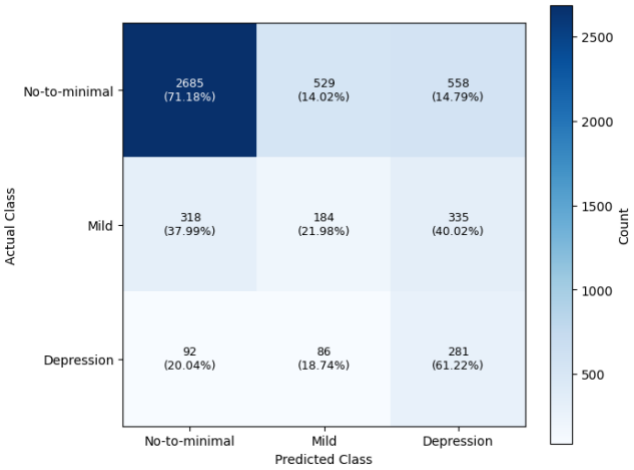

Fig. 3. Confusion matrix of Logistic Regression with the Final 15 FE Only scenario, showing prediction counts and normalized percentages by actual class.

The confusion matrix was used to examine the classification error patterns of Logistic Regression with the Final 15 FE-only scenario. To provide clearer interpretation across imbalanced class sizes, the confusion matrix is presented using both raw counts and normalized percentages, as shown in Fig. 3. Based on Fig. 3, among the 459 respondents in the depression class, 281 respondents were correctly classified as depression, corresponding to 61.22% of the actual

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

depression class. Meanwhile, 92 respondents, or 20.04%, were misclassified as no-to-minimal depression, and 86 respondents, or 18.74%, were misclassified as mild depression. These results indicate that the model was able to identify more than half of the respondents in the depression class, although a considerable proportion of depression cases were still assigned to lower depression categories.

The confusion matrix also shows that 558 respondents from the no-to-minimal depression class and 335 respondents from the mild depression class were predicted as depression. This error pattern explains the low precision for the depression class reported in Table 3. In other words, although the model had a relatively high recall for the depression class, it also produced many false positive predictions. Therefore, the depression-class result should be interpreted as higher sensitivity rather than reliable diagnostic precision. These findings support the use of the model for exploratory and population-level screening support, where sensitivity toward possible depression cases can be useful. However, the high number of false positive predictions indicates that the model should not be used as a stand-alone diagnostic tool. Any individual-level interpretation should require further clinical screening and professional evaluation.

D. Baseline Model Comparison

To examine whether the engineered features were useful beyond Logistic Regression, baseline comparisons were conducted using Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost. All models were evaluated under the same three feature scenarios: Raw Features, Raw + Final 15 FE, and Final 15 FE Only. The comparison focused on two key metrics, namely macro F1-score and recall for the depression class, as summarized in Table 6.

Based on Table 6, the highest macro F1-score across all tested models and feature combinations was achieved by Gradient Boosting with the Raw + Final 15 FE scenario, reaching 0.4518 ± 0.0068. This value was only slightly higher than Logistic Regression with the Final 15 FE Only scenario, which achieved a macro F1-score of 0.4501 ± 0.0104. However, Logistic Regression with the Final 15 FE Only scenario produced the highest recall for the depression class, reaching 0.6122 ± 0.0622.

The effect of the engineered features varied across classifiers. In Logistic Regression, the Final 15 FE Only scenario improved both macro F1-score and depression-class recall compared with Raw Features. A similar pattern was observed in Random Forest, although its depression-class recall remained very low across all scenarios. In contrast, SVM, Gradient Boosting, and XGBoost did not show consistent improvement when using the Final 15 FE Only scenario. These results indicate that the proposed engineered features were most beneficial in the Logistic Regression setting, where explicit interaction-based features can help represent cross-domain relationships in an interpretable linear model. Although Gradient Boosting with Raw + Final 15 FE achieved the highest macro F1-score, the difference from Logistic Regression with Final 15 FE Only was very small. Moreover, Logistic Regression provided a more compact and interpretable model with the highest depression-class recall. Therefore, Logistic Regression with Final 15 FE Only was retained as the main model because it was more aligned with the objective of this study, namely, developing a compact and interpretable cross-domain feature representation while improving sensitivity toward the minority depression class.

Table 6. Baseline model comparison across feature scenarios.

Metric	LR	RF	SVM	GB	XGBoost
Raw Macro F1	0.4400 ± 0.0151	0.3397 ± 0.0084	0.4433 ± 0.0170	0.4479 ± 0.0084	0.4476 ± 0.0080
Raw + FE Macro F1	0.4383 ± 0.0104	0.3465 ± 0.0083	0.4484 ± 0.0052	0.4518 ± 0.0068	0.4491 ± 0.0063
Final FE Macro F1	0.4501 ± 0.0104	0.3632 ± 0.0082	0.4316 ± 0.0069	0.4427 ± 0.0106	0.4450 ± 0.0043
Raw Recall Dep.	0.5360 ± 0.0541	0.0523 ± 0.0161	0.3770 ± 0.0771	0.4749 ± 0.0261	0.4836 ± 0.0484
Raw + FE Recall Dep.	0.5230 ± 0.0756	0.0610 ± 0.0111	0.3791 ± 0.0458	0.4793 ± 0.0669	0.4728 ± 0.0344
Final FE Recall Dep.	0.6122 ± 0.0622	0.0763 ± 0.0121	0.4880 ± 0.0378	0.4728 ± 0.0471	0.4663 ± 0.0462

E. Imbalance-Handling Sensitivity Analysis

Because the dataset had an imbalanced class distribution, an additional sensitivity analysis was conducted to compare several imbalance-handling strategies in the main setting, namely, Logistic Regression with the Final 15 FE Only scenario. The compared strategies included no imbalance handling, class_weight = balanced, SMOTE, and random undersampling. The results are shown in Table 7. Based on Table 7, the model without imbalance handling achieved the highest accuracy, with a value of 0.7421. However, its recall for the depression class was only 0.1002, indicating that the model was strongly biased toward the majority class and failed to identify most respondents in the depression category. This finding confirms that accuracy alone is not sufficient for evaluating model performance in an imbalanced depression classification setting. After

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

imbalance handling was applied, the recall and F1-score for the depression class increased substantially. The class_weight = balanced strategy achieved a macro F1-score of 0.4501 and an F1-score for the depression class of 0.3436, which were the highest among the tested imbalance-handling strategies. SMOTE and random undersampling produced comparable depression-class recall values of 0.5991 and 0.6100, respectively, but neither method clearly outperformed class_weight = balanced.

Table 7. Imbalance-handling sensitivity analysis using Logistic Regression with Final 15 FE Only.

Metric	None	Class weight	SMOTE	Under-sampling
Accuracy	**0.7421**	0.6215	0.6196	0.6089
Macro F1	0.3630	**0.4501**	0.4474	0.4455
Recall Dep.	0.1002	**0.6122**	0.5991	0.6100
F1 Dep.	0.1514	**0.3436**	0.3389	0.3391
PR-AUC Dep.	**0.2707**	0.2665	0.2607	0.2663

The class_weight = balanced strategy was retained as the main imbalance-handling approach because it improved sensitivity toward the minority depression class while preserving the original training data distribution. In contrast, SMOTE modifies the training distribution by generating synthetic samples, while random undersampling reduces the number of majority-class samples. Since the main focus of this study was compact feature engineering rather than imbalance-handling optimization, class_weight = balanced was considered a reasonable and interpretable strategy for the main Logistic Regression evaluation. Fold-level Wilcoxon signed-rank testing showed that the improvement of class_weight = balanced over no imbalance handling was not statistically significant at the 0.05 level, although it showed a practical improvement in depression-class recall. Comparisons between class_weight = balanced and the resampling methods also did not show statistically significant differences. Therefore, the imbalance-handling sensitivity analysis supports the use of class_weight = balanced as a practical strategy, while also indicating that the observed differences among imbalance-handling methods should be interpreted cautiously.

IV. Discussion

A. Model Performance and Practical Interpretation

The results showed that Logistic Regression with the Final 15 FE Only scenario identified more than half of respondents in the depression class, although performance was not evenly distributed across the three depression levels. The model achieved an accuracy of 0.6215 ± 0.0091, a macro F1-score of 0.4501 ± 0.0104, a balanced accuracy of 0.5146 ± 0.0179, and a recall for the depression class of 0.6122 ± 0.0622. The accuracy value should be interpreted carefully because the dataset was imbalanced. The no-to-minimal depression class accounted for 74.4% of the data, whereas the depression class accounted for only 9.1%. A model may obtain relatively high accuracy by favoring the majority class, while still failing to detect respondents in the minority depression class. This pattern was observed in the imbalance-handling sensitivity analysis, where the model without imbalance handling achieved the highest accuracy but produced very low recall for the depression class. Therefore, accuracy alone was not sufficient, and macro F1-score, balanced accuracy, and depression-class recall were considered more informative for evaluating model performance in this imbalanced multiclass setting [38][30].

The per-class results showed that the model performed best on the no-to-minimal depression class, with an F1-score of 0.7819 ± 0.0105. This result was consistent with the dominance of this class in the dataset. In contrast, the mild depression class had the lowest F1-score, at 0.2250 ± 0.0345. This finding suggests that mild depression was the most difficult class to distinguish using the available predictors. One possible reason is that mild depression is an intermediate PHQ-9 severity category, so its symptom pattern may overlap with both no-to-minimal depression and depression in the threshold-based PHQ-9 severity classification [42]. Therefore, sleep, physical activity, and demographic variables alone may not provide enough discriminative information to clearly separate mild depression from neighboring classes. For the depression class, the model achieved an F1-score of 0.3436 ± 0.0256, with higher recall than precision. This indicates that the model was relatively sensitive in identifying respondents with depression, but it also produced many false positive predictions. The confusion matrix supports this interpretation, as many respondents from the no-to-minimal and mild depression classes were predicted as depression. Therefore, the model should not be interpreted as a stand-alone diagnostic tool. Its practical value is more appropriate for exploratory analysis or population-level screening support, where identifying potentially at-risk respondents is useful, but further clinical assessment remains necessary.

B. Impact and Contribution of Cross-Domain Feature Engineering

The cross-domain feature engineering approach was designed to provide compact interaction-based representations for Logistic Regression. Since Logistic Regression models predictors additively, interaction features such as sleepiness_poverty, sleep_isolation, age_sleep_interact, income_activity_interact, and income_sleep_interact were explicitly constructed to represent relationships across sleep, physical activity, and demographic domains while preserving interpretability. The evaluation showed that the Final 15 FE Only scenario achieved a macro F1-score of 0.4501 ± 0.0104, compared with 0.4400 ± 0.0151 for Raw Features. The recall for the depression class also increased from 0.5360 ± 0.0541 to 0.6122 ± 0.0622. These results suggest that the engineered features helped the Logistic Regression

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

Table 8. Comparison with related depression prediction studies.

Study	Dataset/Context	Target	Model/Approach	Main Performance	Focus/Position
ShangGuan et al. [16]	NHANES 2005–2020; physically inactive adults	PHQ-9 > 9	6 ML + nomogram	LR AUC = 0.769; validation AUC = 0.736–0.794	Inactive adults
Vu et al. [18]	NHANES 2013–2014	PHQ-9 ≥ 10	ML + SHAP	XGBoost AUC = 0.69; F1 = 0.69	Binary prediction
Dong et al. [19]	NHANES 2011–2016	Depression risk	11 ML models + SHAP	LR/Lasso test AUC = 0.719	Determinant analysis
Lin et al. [20]	NHANES 2005–2018	Depression risk	LASSO + LR + MR	AUC = 0.626; validation AUC = 0.616	Biomarker prediction
Qu et al. [21]	NHANES 2005–2018; veterans	PHQ-9 ≥ 10	DL + ML models	DL AUC = 0.891; F1 = 0.816	Veteran prediction
Li et al. [22]	NHANES 2005–2008, 2015–2016; OSAHS adults	PHQ-9 ≥ 5	LR, LASSO, RF + nomogram	LR AUC = 0.746	Sleep apnea patients
This study	**NHANES 2017–2018**	**Three-class depression level**	**Compact FE + LR**	**Macro F1 = 0.4501; recall = 0.6122**	**Multiclass compact FE**

model become more sensitive toward the depression class. This improvement may occur because the engineered features summarize clinically and behaviorally relevant patterns, such as non-optimal sleep duration, sleep-related problems, sedentary behavior, and interactions between sleep or activity patterns with age and income-related factors.

However, the improvement should be interpreted as modest. The Wilcoxon signed-rank test showed that the improvement of Final 15 FE Only over Raw Features was not statistically significant at the 0.05 level [30]. In addition, PR-AUC for the depression class did not improve compared with Raw Features. This indicates that the engineered features increased depression-class recall, but did not substantially improve the overall precision-recall trade-off. In other words, the model became more sensitive to possible depression cases, but this sensitivity was accompanied by a considerable number of false positive predictions. The Raw + Final 15 FE scenario did not outperform the Final 15 FE Only scenario, suggesting that combining raw and engineered variables may have introduced redundant information. The baseline comparison also showed that the benefit of the engineered features was not consistent across classifiers. Therefore, the proposed feature set is most relevant in the Logistic Regression setting, where compactness, interpretability, and depression-class sensitivity were prioritized over maximizing overall predictive performance using more complex nonlinear models. The relevance of the engineered features is supported by previous findings on the association between sleep-related factors and depressive symptoms [5][6][8][31], physical activity and sedentary behavior with depressive symptoms [13][14], and socioeconomic or social-related factors with depressive symptoms [32][33]. In this study, these domains were combined into interaction-based features to represent the possible joint contribution of sleep, physical activity, and demographic conditions to depression levels.

To clarify the position of this study, Table 8 summarizes several recent NHANES-based depression prediction studies. As shown in Table 8, several previous NHANES-based studies reported higher overall predictive performance than this study. ShangGuan et al. [16] reported an LR AUC of 0.769 and validation AUC values ranging from 0.736 to 0.794 in physically inactive adults, while Vu et al. [18] reported an XGBoost AUC of 0.69 and an F1-score of 0.69 for binary depression prediction. Dong et al. [19] reported a test AUC of 0.719 using LR/Lasso-based models, and Lin et al. [20] reported an AUC of 0.626 with a validation AUC of 0.616 in a biomarker-based depression risk prediction setting. Qu et al. [21] achieved higher performance, with an AUC of 0.891 and an F1-score of 0.816 in a veteran-specific dataset, while Li et al. [22] reported an LR AUC of 0.746 in adults with obstructive sleep apnea hypopnea syndrome. In comparison, this study achieved a macro F1-score of 0.4501 and depression-class recall of 0.6122 in a three-class depression level classification setting.

The lower overall performance in this study is reasonable because most previous studies formulated depression prediction as a binary classification task, such

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

as PHQ-9 ≥ 10 versus non-depression, whereas this study used a three-class depression level classification setting. Binary classification generally has a clearer decision boundary than multiclass severity classification, especially when the intermediate mild depression class overlaps with adjacent categories. In addition, several previous studies used broader predictor sets, including clinical, dietary, biomarker, or population-specific variables, while this study intentionally focused on a compact set of sleep, physical activity, and demographic interaction features. Thus, the main contribution of this study is not to achieve the highest predictive performance compared with previous studies, but to provide a compact and interpretable feature representation for multiclass depression-level classification in imbalanced NHANES survey data. The proposed approach reduced the feature set from 20 raw predictors to 15 engineered features and improved depression-class recall in the main Logistic Regression setting. However, the overall gain remained modest and was not statistically significant, indicating that the proposed features should be interpreted as improving minority-class sensitivity rather than producing a large overall predictive improvement.

C. Limitations, Clinical Implications, and Future Work

This study has several limitations. First, the data were obtained from NHANES 2017–2018, in which several variables were self-reported. This may introduce subjectivity bias, particularly for sleep, physical activity, sedentary behavior, and depressive symptoms, because such measures may be affected by differences in perception, memory, and reporting [9]. Second, the depression labels were constructed from the total PHQ-9 score; therefore, the predicted classes represent levels of depressive symptoms based on a screening questionnaire rather than clinical diagnosis [26][42]. Third, the dataset had a highly imbalanced class distribution, with the depression class representing only 9.1% of the data, which may affect minority-class prediction performance [38].

Another limitation concerns the discriminative ability across adjacent depression levels. The mild depression class remained difficult to identify, as shown by its low F1-score. This suggests that the three-class formulation based on PHQ-9 thresholds may be difficult to separate using only sleep, physical activity, and demographic predictors. In addition, the precision for the depression class remained low at 0.2390 ± 0.0162, although the recall reached 0.6122 ± 0.0622. This indicates that the model was relatively sensitive in identifying possible depression cases but still produced many false positive predictions. Considering these limitations, the proposed model should not be used as a stand-alone diagnostic tool. Instead, its practical value is more appropriate for exploratory and population-level analysis, particularly for examining how compact cross-domain features from sleep, physical activity, and demographic information contribute to depression-level classification in survey data. In public health contexts, this approach may support initial risk stratification or help identify groups that require further assessment. However, any individual-level interpretation should be followed by validated clinical screening and professional evaluation.

Future research can be developed in several directions. Binary and ordinal classification formulations can be compared with the current three-class formulation to examine whether adjacent PHQ-9 severity levels can be modeled more effectively. Threshold adjustment and probability calibration may also be explored to reduce false positive predictions and improve precision for the depression class. In addition, future studies can incorporate broader predictors, such as medical history, medication use, alcohol consumption, social support, and environmental factors. Nonlinear but interpretable machine learning models can also be evaluated using the same engineered features. Finally, the proposed approach should be tested on other NHANES cycles or external mental health datasets to examine its consistency, generalisability, and external validity across populations [43].

V. Conclusion

This study developed a compact cross-domain feature engineering approach for classifying depression levels using sleep, physical activity, and demographic data from NHANES 2017–2018. A total of 5,068 respondents were included after preprocessing and PHQ-9 label construction. The original 20 raw predictors were transformed into 15 engineered features representing sleep patterns, sleep-related problems, physical activity, sedentary behavior, and cross-domain interactions with age and income. The main Logistic Regression model using the Final 15 FE Only scenario achieved an accuracy of 0.6215 ± 0.0091, a macro F1-score of 0.4501 ± 0.0104, a balanced accuracy of 0.5146 ± 0.0179, and a depression-class recall of 0.6122 ± 0.0622. Compared with Raw Features, the Final 15 FE Only scenario improved macro F1-score from 0.4400 ± 0.0151 to 0.4501 ± 0.0104 and depression-class recall from 0.5360 ± 0.0541 to 0.6122 ± 0.0622. However, the Wilcoxon signed-rank test indicated that these improvements were not statistically significant at the 0.05 level. The findings suggest that the proposed engineered features improved sensitivity toward the depression class in the Logistic Regression setting, but the overall predictive gain remained modest. The model still showed limited precision for the depression class and low performance for the mild depression class, indicating that multiclass depression-level classification remains challenging in imbalanced survey data. Therefore, the model should not be used as a stand-alone diagnostic tool. Its practical value is more suitable for exploratory analysis and population-level screening support. Future work should test the proposed approach on other NHANES cycles or external mental health datasets and explore alternative classification formulations, probability calibration, threshold adjustment, and broader predictors to improve the separation of adjacent depression levels.

References

[1] J. Liu, Y. Liu, W. Ma, Y. Tong, and J. Zheng,

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

"Temporal and spatial trend analysis of all-cause depression burden based on Global Burden of Disease (GBD) 2019 study," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-62381-9.

[2] World Health Organization, "Depressive disorder (depression)," 2025, *World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] Y. Sun, Z. Kong, Y. Song, J. Liu, and X. Wang, "The validity and reliability of the PHQ-9 on screening of depression in neurology: a cross sectional study," *BMC Psychiatry*, vol. 22, no. 1, pp. 1–12, 2022, doi: 10.1186/s12888-021-03661-w.

[4] M. A. Rahman, T. A. Dhira, A. R. Sarker, and J. Mehareen, "Validity and reliability of the Patient Health Questionnaire scale (PHQ-9) among university students of Bangladesh," *PLoS One*, vol. 17, no. 6 June, pp. 1–13, 2022, doi: 10.1371/journal.pone.0269634.

[5] Y. J. Um *et al.*, "Association of changes in sleep duration and quality with incidence of depression: A cohort study," *J. Affect. Disord.*, vol. 328, no. August 2022, pp. 64–71, 2023, doi: 10.1016/j.jad.2023.02.031.

[6] H. J. Joo, K. A. Kwon, J. Shin, S. Park, and S. I. Jang, "Association between sleep quality and depressive symptoms," *J. Affect. Disord.*, vol. 310, no. February, pp. 258–265, 2022, doi: 10.1016/j.jad.2022.05.004.

[7] K. Joshi, M. J. Cambron-Mellott, H. Costantino, A. Pfau, and M. K. Jha, "The real-world burden of adults with major depressive disorder with moderate or severe insomnia symptoms in the United States," *J. Affect. Disord.*, vol. 323, no. September 2022, pp. 698–706, 2023, doi: 10.1016/j.jad.2022.12.005.

[8] S. Wang, M. E. Rossheim, R. R. Nandy, and U. S. Nguyen, "Interaction between sleep duration and trouble sleeping on depressive symptoms among U.S. adults, NHANES 2015-2018," *J. Affect. Disord.*, vol. 351, no. July 2023, pp. 285–292, 2024, doi: 10.1016/j.jad.2024.01.260.

[9] V. L. Amelia, H. J. Jen, T. Y. Lee, L. F. Chang, and M. H. Chung, "Comparison of the Associations between Self-Reported Sleep Quality and Sleep Duration Concerning the Risk of Depression: A Nationwide Population-Based Study in Indonesia," *Int. J. Environ. Res. Public Health*, vol. 19, no. 21, 2022, doi: 10.3390/ijerph192114273.

[10] M. Rovero *et al.*, "Subtypes of major depressive disorders and objectively measured physical activity and sedentary behaviors in the community," *Compr. Psychiatry*, vol. 129, no. December 2023, p. 152442, 2024, doi: 10.1016/j.comppsych.2023.152442.

[11] C. H. Wang and N. Peiper, "Association Between Physical Activity and Sedentary Behavior With Depressive Symptoms Among US High School Students, 2019," *Prev. Chronic Dis.*, vol. 19, no. 1, pp. 1–13, 2022, doi: 10.5888/PCD19.220003.

[12] V. Gianfredi *et al.*, "Daily patterns of physical activity, sedentary behavior, and prevalent and incident depression—The Maastricht Study," *Scand. J. Med. Sci. Sports*, vol. 32, no. 12, pp. 1768–1780, 2022, doi: 10.1111/sms.14235.

[13] Y. Guo, K. Li, Y. Zhao, C. Wang, H. Mo, and Y. Li, "Association between long-term sedentary behavior and depressive symptoms in U.S. adults," *Sci. Rep.*, vol. 14, no. 1, pp. 1–9, 2024, doi: 10.1038/s41598-024-55898-6.

[14] Y. Meng *et al.*, "The association of physical activity and sedentary behavior with depression in US adults: NHANES 2007–2018," *Front. Public Heal.*, vol. 12, no. June, pp. 1–11, 2024, doi: 10.3389/fpubh.2024.1404407.

[15] D. Colledani, P. Anselmi, and E. Robusto, "Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major depressive disorder," *Psychiatry Res.*, vol. 322, no. October 2022, p. 115127, 2023, doi: 10.1016/j.psychres.2023.115127.

[16] Y. ShangGuan *et al.*, "Development and validation of a machine learning model for predicting the risk of current depression in physically inactive adults in the United States," *J. Affect. Disord.*, vol. 394, p. 120555, 2026, doi: 10.1016/j.jad.2025.120555.

[17] T. Salahudeen, M. Maalouf, I. M. Elfadel, and H. F. Jelinek, "Predicting depression severity using machine learning models: Insights from mitochondrial peptides and clinical factors," *PLoS One*, vol. 20, no. 5, pp. 1–29, 2025, doi: 10.1371/journal.pone.0320955.

[18] T. Vu *et al.*, "Prediction of depressive disorder using machine learning approaches: findings from the NHANES," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, pp. 1–12, 2025, doi: 10.1186/s12911-025-02903-1.

[19] Y. Dong, H. Wen, C. Lu, J. Li, and Q. Zheng, "Predicting depression risk with machine learning models: identifying familial, personal, and dietary determinants," *BMC Psychiatry*, vol. 25, no. 1, 2025, doi: 10.1186/s12888-025-07182-8.

[20] L. Lin, L. Zhang, J. Zhang, and D. Ding, "A Novel Depression Risk Prediction Model Using NHANES Data With Mendelian Randomization Validation," *Brain Behav.*, vol. 15, no. 7, pp. 1–15, 2025, doi: 10.1002/brb3.70674.

[21] Z. Qu *et al.*, "Identifying depression in the United States veterans using deep learning algorithms, NHANES 2005–2018," *BMC Psychiatry*, vol. 23, no. 1, pp. 1–10, 2023, doi: 10.1186/s12888-023-05109-9.

[22] E. Li, F. Ai, and C. Liang, "A machine learning model to predict the risk of depression in US adults with obstructive sleep apnea hypopnea syndrome: a cross-sectional study," *Front. Public Heal.*, vol.

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

11, no. January, 2023, doi: 10.3389/fpubh.2023.1348803.

[23] L. Yang *et al.*, "Application of machine learning in depression risk prediction for connective tissue diseases," *Sci. Rep.*, vol. 15, no. 1, pp. 1–10, 2025, doi: 10.1038/s41598-025-85890-7.

[24] Centers for Disease Control and Prevention, "2017–2018 Questionnaire Data - Continuous NHANES," 2020, *National Center for Health Statistics*. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2018.

[25] K. Kroenke, "PHQ-9: global uptake of a depression scale," *World Psychiatry*, vol. 20, no. 1, pp. 135–136, 2021, doi: 10.1002/wps.20821.

[26] Z. F. Negeri *et al.*, "Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression : updated systematic review and individual participant data meta-analysis," *BMJ*, vol. 374, p. n2183, 2021, doi: 10.1136/bmj.n2183.

[27] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, 2025, doi: 10.3390/ai6100257.

[28] K. Psychogyios, L. Ilias, C. Ntanos, and D. Askounis, "Missing Value Imputation Methods for Electronic Health Records," *IEEE Access*, vol. 11, no. March, pp. 21562–21574, 2023, doi: 10.1109/ACCESS.2023.3251919.

[29] W. Ren, Z. Liu, Y. Wu, Z. Zhang, S. Hong, and H. Liu, "Moving Beyond Medical Statistics : A Systematic Review on Missing Data Handling in Electronic Health Records," *Heal. Data Sci.*, vol. 4, pp. 1–19, 2024, doi: 10.34133/hds.0176.

[30] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[31] L. Chunnan, S. Shaomei, and L. Wannian, "The association between sleep and depressive symptoms in US adults: data from the NHANES (2007-2014)," *Epidemiol. Psychiatr. Sci.*, vol. 31, 2022, doi: 10.1017/S2045796022000452.

[32] H. Zare, N. S. Meyerson, C. A. Nwankwo, and R. J. Thorpe, "How Income and Income Inequality Drive Depressive Symptoms in U.S. Adults, Does Sex Matter: 2005–2016," *Int. J. Environ. Res. Public Health*, vol. 19, no. 10, 2022, doi: 10.3390/ijerph19106227.

[33] S. Kim, Y. S. Jang, and E. C. Park, "Associations between social isolation, withdrawal, and depressive symptoms in young adults: a cross-sectional study," *BMC Psychiatry*, vol. 25, no. 1, 2025, doi: 10.1186/s12888-025-06792-6.

[34] Y. Jiang, M. Zhang, and J. Cui, "The relationship between sedentary behavior and depression in older adults: A systematic review and meta-analysis," *J. Affect. Disord.*, vol. 362, no. July, pp. 723–730, 2024, doi: 10.1016/j.jad.2024.07.097.

[35] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, 2024, doi: 10.1136/bmj-2023-078378.

[36] Y. Hu *et al.*, "Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality," *J. Med. Internet Res.*, vol. 27, pp. 1–6, 2025, doi: 10.2196/77721.

[37] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, 2022, doi: 10.1016/j.cmpb.2022.107161.

[38] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, 2024, doi: 10.1007/s10462-024-10759-6.

[39] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23042333.

[40] L. Sasse *et al.*, "Overview of leakage scenarios in supervised machine learning," *J. Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01193-8.

[41] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Heal.*, vol. 6, no. July, 2024, doi: 10.3389/fdgth.2024.1430245.

[42] B. Levis *et al.*, "Data-Driven Cutoff Selection for the Patient Health Questionnaire-9 Depression Screening Tool," *JAMA Netw. open.*, vol. 7, no. 11, p. e2429630, 2024, doi: 10.1001/jamanetworkopen.2024.29630.

[43] G. S. Collins *et al.*, "Evaluation of clinical prediction models (part 1): from development to external validation," *BMJ*, no. part 1, 2024, doi: 10.1136/bmj-2023-074819.

Author Biography

Nila Yoga Tama Nurwati has been a student in the Computer Science Program at Lambung Mangkurat University since 2022. Her academic interests focus on Data Science, particularly in data analysis, machine learning, and predictive modeling. She is passionate about exploring innovative approaches to solving real-world problems through data-driven methods. In addition to her academic activities, she actively participates in collaborative research projects aimed at enhancing her practical skills

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356

and expanding her knowledge in the field of Data Science. Through continuous learning and professional development, she aspires to contribute to technological advancement, especially in developing effective and efficient data-based solutions. She is also committed to innovation and interdisciplinary collaboration, believing that data has significant potential to create positive impacts across various sectors and aspects of life. She can be contacted via email at nla97742@gmail.com.

Fatma Indriani is a lecturer in the Department of Computer Science at Lambung Mangkurat University, with a strong research interest in Data Science. Before pursuing an academic career, she completed her undergraduate studies in the Informatics Department at the Bandung Institute of Technology. In 2008, she began her journey as a lecturer at Lambung Mangkurat University, contributing to the field of Computer Science through teaching and research. To further expand her expertise, she pursued a master's degree at Monash University, Australia, which she successfully completed in 2012. Her academic journey continued with a doctorate in Bioinformatics from Kanazawa University, Japan, which she completed in 2022. With a focus on both Data Science and Bioinformatics, she actively engages in research, exploring innovative ways to leverage data-driven technologies for scientific advancement. Her dedication to academia and research allows her to contribute significantly to the development of knowledge in her field, while also mentoring students and collaborating on interdisciplinary projects. She can be contacted via email at f.indriani@ulm.ac.id

Friska Abadi earned his Bachelor's degree in Computer Science from Lambung Mangkurat University in 2011, and subsequently completed his Master's degree in Informatics at STMIK AMIKOM Yogyakarta in 2016. He is currently serving as a lecturer in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University. His academic and research interests focus on data mining and software engineering, particularly in the application of computational methods for data analysis and software quality improvement. In addition to his teaching responsibilities, he is actively involved in academic supervision and research activities within the department. He has contributed to various academic publications and participated in technology-related research and community service programs. He can be contacted via email at friska.abadi@ulm.ac.id.

Dodon Turianto Nugrahadi is a lecturer in the Department of Computer Science, Lambung Mangkurat University. He received his bachelor's degree in Informatics Engineering from UK Petra, Surabaya, in 2004 and his master's degree in Informatics Engineering from Gadjah Mada University, Yogyakarta, in 2009. His research interests include Data Science, Computer Networks, Internet of Things (IoT), and Quality of Service (QoS). He has been involved in teaching, academic supervision, and research activities related to computer networks, data-based systems, and applied computing. His current research focuses on the use of data science and network technologies to support computational problem-solving in various application domains. He can be contacted via email at dodonturianto@ulm.ac.id.

Rudy Herteno obtained his bachelor's degree in Computer Science from Lambung Mangkurat University in 2011. After completing his undergraduate studies, he pursued a career as a software developer for several years, gaining extensive experience in designing and implementing software systems. During his professional career, he contributed to the development of various software applications, particularly those aimed at supporting administrative and operational activities within local government institutions. In 2017, he earned his master's degree in Informatics from STMIK Amikom University. He is currently serving as a lecturer in the Computer Science Program at Lambung Mangkurat University. His research interests include software engineering, software defect prediction, and deep learning, with a focus on improving software quality, optimizing error detection mechanisms, and developing artificial intelligence-based solutions. He can be contacted via email at rudy.herteno@ulm.ac.id.

Corresponding author: Fatma Indriani, f.indriani@ulm.ac.id, Department of Computer Science, Faculty of Mathematics and Natural Science, Banjarbaru, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.356