Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 27 NO 06 2026, pp 1162-1177
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol27-iss06/774

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Hypertension Prediction by Comparing AMFormer, XGBoost, and Logistic Regression

Article Info

Article history :
Received October 12, 2026
Revised November 10, 2026
Accepted November 17, 2026
Published December 30, 2026
In Press

Keywords :
Hypertension prediction,
machine learning,
AMFormer,
IFLS,
SHAP

Eka Dela Sari[1], A'yunin Sofro[2*]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Surabaya, Surabaya, Indonesia
[2]Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, Indonesia

Abstract. Early identification of individuals with hypertension using routinely available characteristics may support population-level screening, particularly where direct clinical assessment is limited. However, hypertension remains a substantial health concern, and the predictive benefit of increasingly complex models over conventional approaches remains uncertain. This study compared Logistic Regression, XGBoost, AMFormer Full, and AMFormer Ablation for hypertension classification using the fifth wave of the Indonesian Family Life Survey (IFLS5). The analytical dataset comprised 32,413 respondents, of whom 28.85% were classified as hypertensive. Hypertension status was defined using a composite of measured blood pressure, self-reported physician diagnosis, and antihypertensive medication use. Fourteen demographic, anthropometric, socioeconomic, behavioral, physical-activity, and metabolic predictors were used. Model performance was evaluated on an independent test set and through repeated stratified cross-validation, while Shapley Additive exPlanations (SHAP) was applied to interpret AMFormer Full. AMFormer Ablation achieved the highest test ROC-AUC (0.7992) and average precision (0.6242), exceeding the positive-class prevalence baseline (0.2886). However, performance differences among the four models were small. Cross-validation also showed nearly identical performance between AMFormer variants. SHAP identified age, BMI, weight, sex, and education level as the five most influential predictors. Overall, AMFormer was competitive but showed no substantial predictive advantage over conventional models, and further external validation is required.

This is an open acces article under the CC-BY license.

This is an open access article distributed under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2026 by author.

Corresponding Author :
A'yunin Sofro
Department of Actuarial Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Surabaya, Surabaya, Indonesia
Email : ayuninsofro@unesa.ac.id

1. Introduction

Hypertension is a major public health concern because uncontrolled blood pressure increases the risk of cardiovascular disease, stroke, and kidney complications. Early identification supports preventive strategies and timely health assessment. The World Health Organization highlights the global burden of hypertension and the need to improve its detection and control [1]. Predictive modelling offers an approach to identifying individuals who may require further assessment using routinely collected health information. Previous systematic reviews have examined statistical and machine-learning methods for hypertension prediction, reporting varying performance across populations and study designs [2-3].

Machine-learning-based hypertension screening has also been investigated in Indonesia. Estiko et al. [4] evaluated models using non-laboratory risk factors, while Septian et al. [5] developed personalized hypertension prediction models using Indonesian health data. Alfaqeeh et al. [6] examined factors associated with hypertension among Indonesian adults using IFLS data. Other studies have applied machine-learning methods to health-related classification and biomarker analysis [7-8]. However, differences in data sources, predictors, outcome definitions, and modelling strategies limit direct comparisons, highlighting the need to evaluate alternative architectures under consistent experimental procedures.

Hypertension is associated with demographic, anthropometric, socioeconomic, behavioral, and metabolic characteristics [2-3], [6]. Accordingly, this study uses 14 predictors covering age, sex, anthropometric measures, socioeconomic characteristics, smoking status, physical activity, diabetes, and cholesterol status. Direct measures of dietary salt intake, alcohol consumption, and family history of hypertension were not identified in the IFLS5 questionnaire sections examined. Although parental causes of death were available, they did not directly establish a family history of hypertension. These limitations should be considered when interpreting the findings.

Despite their increasing use, deep learning models do not consistently outperform conventional machine-learning methods on structured tabular data, partly because heterogeneous numerical and categorical features present challenges for representation learning [9]. Recent research has also applied gradient-boosting algorithms to classification tasks using survey-based data [10]. The Arithmetic-Modified Transformer (AMFormer) incorporates explicit arithmetic feature interactions to represent relationships among tabular features [11]. In health-related data, the contribution of one characteristic may depend on another; for example, the predictive contribution of BMI may vary across age groups. Multiplicative interactions provide a mechanism for representing such combined feature patterns. Previous research has evaluated AMFormer for new-onset hypertension prediction using the China Health and Nutrition Survey [12]. However, its performance on Indonesian population data and the contribution of its multiplicative interaction component remain to be investigated.

This study uses individual-level data from the fifth wave of the Indonesia Family Life Survey (IFLS5), collected in 2014–2015. Despite its age, the dataset provides comprehensive demographic, socioeconomic, and health information for examining population-level relationships, although the findings should be interpreted within the survey period rather than as representations of current hypertension patterns. Previous studies have used IFLS data to investigate multimorbidity and other health-related characteristics [13], while its longitudinal design and population coverage are documented in the survey guidelines [14]. This study compares Logistic Regression, XGBoost, AMFormer Full, and AMFormer Ablation to evaluate predictive performance and examine feature contributions using SHAP.

2. Experimental Section

2.1. Data Source and Analytical Dataset

This study used individual-level data from the fifth wave of the Indonesia Family Life Survey (IFLS5), a longitudinal household survey collecting demographic, socioeconomic, and health information in Indonesia. Conducted from late 2014 to early 2015, IFLS5 covered 50,148 individuals from 16,204 households across 13 provinces [14].

The original data were processed to construct an analytical dataset for hypertension classification. Respondents were included when the information required to determine hypertension status and the selected predictors was available. Individuals with insufficient outcome information were excluded, resulting in a final analytical dataset of 32,413 unique respondents.

2.2. Response Variable and Predictor Variables

The analytical dataset consisted of one binary response variable and 14 predictors. Hypertension status was defined using three components: measured blood pressure, self-reported physician diagnosis, and antihypertensive medication use. Respondents were classified as hypertensive if at least one component was positive. Non-hypertensive status was assigned only when all three components were available and negative; respondents with incomplete or unknown outcome information and no positive indicator were excluded.

The predictors represented demographic, anthropometric, socioeconomic, behavioral, physical-activity, and metabolic characteristics: age, sex, marital status, education level, main activity, height, weight, BMI, smoking status, vigorous activity, moderate activity, walking activity, diabetes, and cholesterol status. These comprised four continuous and ten categorical or binary variables. Continuous variables were standardized, while categorical variables were encoded according to their defined categories. Missing diabetes and cholesterol values were retained as explicit categories. Variable definitions and coding are provided in Table 1.

Table 1. Description and coding of the variables used in the study

Variable	Description and Categories	Type
Hypertension	Composite Hypertension status (No, Yes)	Binary
Age	Age of respondent, measured in years	Continuous
Sex	Sex of respondent (Male, Female)	Categorical
Marital status	Unmarried, Married, Separated, Divorced, Widow, Cohabiting	Categorical
Education level	Primary, Lower secondary, Upper secondary, Higher, Other	Categorical
Main activity	Working, Job searching, School, Housekeeping, Retired, Sick/disabled, Other	Categorical
Height	Height of respondent	Continuous
Weight	Weight of respondent	Continuous
BMI	Body mass index calculated from height and weight	Continuous
Smoking status	Smoking status (Never, Former, Current)	Categorical
Vigorous active	Vigorous physical activity (No, Yes)	Binary
Moderate active	Moderate physical activity (No, Yes)	Binary
Walking active	Walking activity (No, Yes)	Binary
Diabetes	Reported diabetes status (Yes, Missing)	Binary
Cholesterol	Reported cholesterol status (Yes, Missing)	Binary

2.3. Data Preprocessing and Experimental Design

The analytical dataset was divided into training, validation, and independent test sets using stratified sampling with a random seed of 42. The resulting partitions contained 22,689, 4,862, and 4,862 observations, respectively. Stratification preserved the outcome proportions across partitions, while the test set was reserved for final evaluation [15-16].

Preprocessing parameters were derived exclusively from training data. Continuous variables were median-imputed and standardized, while categorical variables were imputed and encoded using the defined procedures. Missing diabetes and cholesterol values were retained as explicit categories. During cross-validation, preprocessing was fitted separately within each training fold and applied to its corresponding validation fold to prevent information leakage [17-18].

The training set was used for model development and repeated stratified three-fold cross-validation with three repetitions, producing nine validation splits. The validation set supported neural-network checkpoint selection based on ROC-AUC, while the independent test set remained excluded from training, cross-validation, and model selection. The same partitions and preprocessing procedures were applied across all models to ensure a consistent comparison.

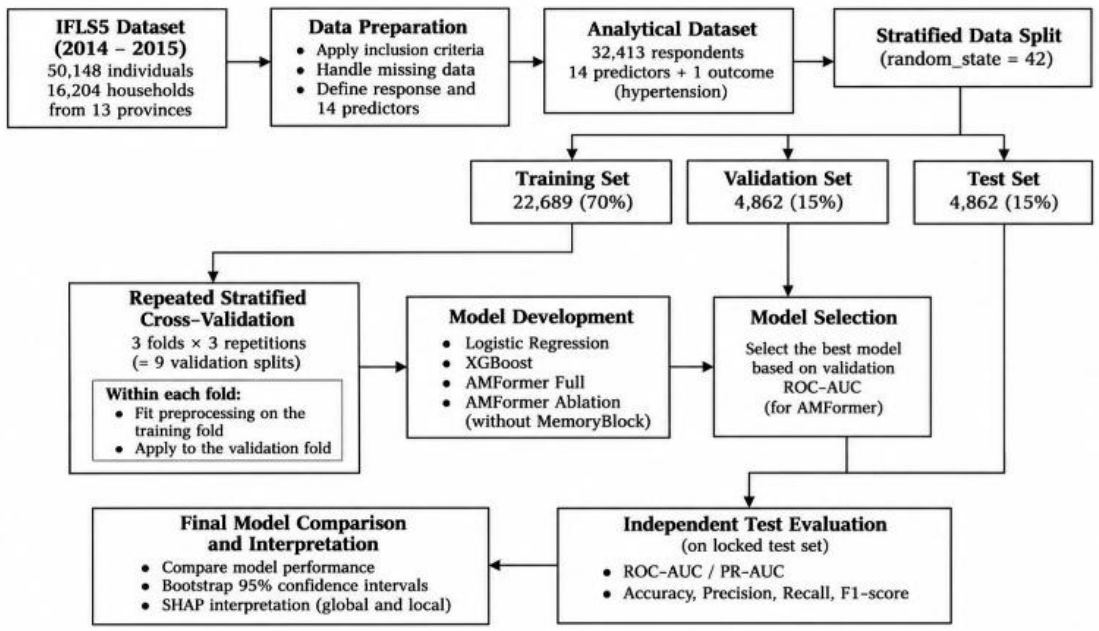


Figure 1. Research workflow

2.4. Model Development

2.4.1. Logistic Regression

Logistic Regression was used as a statistical baseline for binary hypertension classification. It estimates the probability of a positive outcome by applying the logistic function to a linear combination of predictor variables. Its established use in clinical prediction and straightforward formulation make it a suitable reference for evaluating more complex models [19-21]. For an observation i, the predicted probability of belonging to the positive class is expressed as

$$\hat{p}_i = \sigma\left(\beta_0 + x_i^T \beta\right) \tag{1}$$

where x_i represents the predictor vector, β_0 is the intercept, and β is the vector of regression coefficients. The logistic sigmoid function is defined as

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

The sigmoid function transforms the linear predictor into a probability ranging from 0 to 1, allowing the model to estimate the probability of hypertension for each respondent [21-22].

The Logistic Regression model was implemented using the `lbfgs` solver with a maximum of 2,000 iterations. The random_state parameter was set to 42, while the class-weight parameter was left at its default setting (None). Therefore, no explicit class weighting was applied during model fitting. The model was trained using the preprocessed training data and evaluated on the independent test set.

2.4.2. XGBoost

XGBoost is an ensemble algorithm based on gradient-boosted decision trees, in which successive trees improve existing predictions. Its ability to capture nonlinear relationships and interactions among predictors makes it a relevant approach for structured prediction problems, including health-related applications [2, 3, 9]. Recent empirical research has also evaluated XGBoost alongside conventional machine learning and deep learning models across real-world clinical datasets, demonstrating the relevance of gradient-boosted tree methods for clinical risk prediction [23]. The algorithm incorporates regularization to control model complexity and improve generalization [9]. For an observation i, the prediction of an XGBoost model is expressed as the sum of the outputs of K regression trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(x_i) \tag{3}$$

where f_k represents the k-th regression tree and K denotes the number of boosting iterations. The corresponding regularized objective function is

$$\mathcal{L} = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

where $l(y_i, \hat{y}_i)$ is the loss function measuring prediction error and $\Omega(f_k)$ represents the regularization term for the tree structure. The objective balances prediction loss against model complexity through regularization [9].

XGBoost was configured with 300 estimators, maximum tree depth of 4, learning rate of 0.05, subsample ratio of 0.8, and column subsampling ratio of 0.8. The `binary:logistic` objective and `log-loss` evaluation metric were used, with random seed 42 and all available CPU threads. Hyperparameters were specified manually without systematic hyperparameter search.

2.4.3. AMFormer

The Arithmetic-Modified Transformer (AMFormer) is a Transformer-based architecture for tabular data that represents variables as feature tokens and models their relationships through attention mechanisms [9], [24]. It extends standard attention by incorporating explicit additive and multiplicative feature interactions [11]. In the standard Transformer attention mechanism, the input feature-token representation matrix X is projected into query, key, and value matrices using learnable weight matrices:

$$Q = XW^{Q}, \qquad K = XW^{K}, \qquad V = XW^{V} \tag{5}$$

where W^Q, W^K, and W^v are learnable projection matrices. The standard attention operation is then expressed as

$$O^A = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{6}$$

where d is the dimensionality of the query and key representations. The AMFormer architecture modifies this attention framework by introducing arithmetic feature interactions and sparse attention operations [7]. The arithmetic component can be understood through two types of operations. For a group of feature representations e_j, an additive interaction can be represented as

$$S_g = \sum_{j=1}^{m_g} e_j \tag{7}$$

where as a multiplicative interaction can be represented conceptually as

$$P_g = \prod_{j=1}^{n_g} e_j \tag{8}$$

where m_g and n_g denote the numbers of features involved in the corresponding groups. These expressions clarify the distinction between additive and multiplicative feature interactions introduced by AMFormer. The original AMFormer implementation controls the number of features participating in these operations through `sum_num_per_group` and `prod_num_per_group`, with the latter corresponding to multiplicative feature interactions [11]. For the multiplicative pathway, the feature representation is transformed using

$$X_{log} = \log(ReLU(X) + \varepsilon) \tag{9}$$

where ε is a small positive constant. The transformed representation is subsequently processed through the multiplicative attention mechanism and an exponential operation. In the original AMFormer architecture, additive and multiplicative feature interactions are incorporated into the model through its arithmetic interaction mechanism [11].

The AMFormer configuration used an embedding dimension of 192, three Transformer layers, eight attention heads, attention dropout of 0.2, and feed-forward dropout of 0.1. A classification token was used for the final prediction. The arithmetic configuration consisted of `groups=[14,14,14]`, `sum_num_per_group=[8,6,4]`, and `prod_num_per_group=[4,3,3]`, with `cluster=True` and `target_mode="mix"`. The multiplicative interaction module was enabled using `use_prod=True`. In addition, `token_descent=False` was specified in the implementation. Therefore, the full model used in this study combined the multiplicative MemoryBlock with standard attention and feed-forward processing, rather than activating separate additive and multiplicative MemoryBlock pathways simultaneously.

The architectural hyperparameters were specified manually rather than optimized through an automated search procedure such as grid search, random search, or Bayesian optimization. During model training, checkpoint selection was based on validation ROC-AUC, and the selected checkpoint was subsequently used for independent test evaluation. Thus, validation-based checkpoint selection should be distinguished from systematic hyperparameter optimization.

2.4.4. AMFormer Ablation

Ablation analysis was used to determine the contribution of the multiplicative interaction mechanism within AMFormer. In an ablation experiment, a specific component of an architecture is removed while the remaining modelling conditions are kept consistent, allowing its contribution to be examined independently. Such controlled comparisons are useful when assessing whether a particular architectural mechanism provides additional predictive value beyond the underlying model [9], [11]. In the full AMFormer configuration used in this study, the multiplicative MemoryBlock is combined with standard attention and feed-forward processing. Its layer structure can be represented as

$$AMFormer_{Full} = Multiplicatve\ MemoryBlok + Standard\ Attention + FeedForward\ Module \quad (10)$$

In the ablated architecture, the multiplicative MemoryBlock is replaced with an identity operation, while standard attention and feed-forward processing are retained:

$$AMFormer_{Ablation} = Identity\ Operation + Standard\ Attention + FeedForward\ Module \quad (11)$$

The identity operation passes the input representation to the subsequent module without applying the multiplicative interaction transformation. In the implementation, the full model uses `use_prod=True`, whereas the ablation model uses `use_prod=False`. This parameter determines whether the multiplicative MemoryBlock is included in the Transformer layers [11].

All other AMFormer settings were kept unchanged between the two models, including the feature representation, embedding dimension, number of Transformer layers and attention heads, dropout settings, arithmetic group configuration, data partitions, preprocessing procedures, and training strategy. Consequently, the comparison was designed to isolate the contribution of the multiplicative MemoryBlock rather than compare two independently configured architectures.

The ablation should therefore be interpreted as AMFormer without the multiplicative MemoryBlock, while retaining standard attention and feed-forward processing. It should not be described as an architecture that retains an additive MemoryBlock pathway, because the implementation uses standard attention when `token_descent=False`. This distinction is important for accurately interpreting the contribution of the multiplicative interaction mechanism in the experimental results.

2.5. Model Evaluation and Statistical Analysis

Model performance was evaluated using ROC-AUC and average precision (AP) for discrimination, alongside accuracy, precision, recall, specificity, and F1-score at the selected decision thresholds. These metrics provide complementary assessments of predictive performance [25-26].

ROC-AUC was used to evaluate the ability of each model to discriminate between respondents with and without hypertension across different decision thresholds. The receiver operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR). Conceptually, ROC-AUC is expressed as

$$ROC - AUC = \int_0^1 TPR(u)du, \qquad u = FPR \quad (12)$$

where

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN} \quad (13)$$

Here, (TP), (TN), (FP), and (FN) denote true positives, true negatives, false positives, and false negatives, respectively. In this study, ROC-AUC was calculated from predicted probabilities using the `roc_auc_score` function. This metric evaluates discrimination across thresholds rather than the correctness of class labels at a single threshold [25-26]. Precision–recall performance was summarized using average precision (AP), calculated with the `average_precision_score` function in the analysis notebook. AP is defined as

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{14}$$

where P_n and R_n denote precision and recall at threshold (n), respectively. This formulation weights precision by the change in recall between successive thresholds. AP is not identical to the area obtained by applying trapezoidal integration to a precision–recall curve [25-26]. Accordingly, the metric reported as PR-AUC in the experimental results represents AP as implemented in the notebook.

Threshold-dependent metrics were calculated using thresholds of 0.30 for Logistic Regression, 0.30 for XGBoost, 0.50 for AMFormer Full, and 0.50 for AMFormer Ablation. Accuracy, precision, recall, specificity, and F1-score were calculated using their standard definitions [25-26]. These metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN}, \quad Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{15}$$

$$Specificity = \frac{TN}{TN + FP}, \qquad F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

Final performance was evaluated on the independent test set. The 95% confidence intervals for ROC-AUC and AP were estimated using 5,000 percentile bootstrap resamples with random seed 42. Resamples containing only one outcome class were excluded, and the 2.5th and 97.5th percentiles were used as confidence limits [19], [16], [26].

AMFormer Full and AMFormer Ablation were compared using paired two-sided Wilcoxon signed-rank tests on ROC-AUC and AP across the nine cross-validation splits, with a significance level of 0.05. Because repeated cross-validation may produce dependent estimates through overlapping training subsets, these comparisons were treated as exploratory [27-28].

2.6. Model Interpretation

Shapley Additive exPlanations (SHAP) was used to interpret predictions from AMFormer Full. Based on Shapley values, SHAP attributes model outputs to individual features and supports local explanations of individual predictions as well as global summaries across observations [29-30]. For a model (f) and an observation (x), the SHAP value of feature (j) is defined as

$$\phi_j(x) = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{j\}) - f_x(S)] \tag{17}$$

where (F) denotes the complete set of predictor features, (S) represents a subset of features that excludes feature (j), and ($f_x(S)$) denotes the model output associated with subset (S) under the SHAP feature-masking formulation. The term ($f_x(S \cup \{j\}) - f_x(S)$) represents the marginal contribution of

feature (j) when added to subset (S). The weighting factor accounts for the possible subset sizes and orders in which the feature can be introduced [29], [31].

In this study, SHAP values were estimated using the `PermutationExplainer` implementation. A background sample of 200 training observations and an explanation sample of 1,000 independent test observations were selected using random seed 42. The maximum number of model evaluations was set to 29, corresponding to $(2p + 1)$, where $(p = 14)$ is the number of predictor features. The explainer estimates feature attributions through permutations rather than explicitly enumerating all feature subsets [29-30]. Global feature importance was quantified using the mean absolute SHAP value for each feature:

$$I_j = \frac{1}{N}\sum_{i=1}^{N}\left|\phi_j(x_j)\right| \tag{18}$$

where I_j denotes the global importance of feature (j), (N) is the number of observations in the explanation sample, and $\phi_j(x_j)$ is the SHAP value of feature (j) for observation (i). Larger values of (I_j) indicate greater average attribution magnitude across the analyzed observations, irrespective of whether the contributions are positive or negative [29], [31].

SHAP summary plots were used to examine the distribution and direction of feature contributions, while global importance plots ranked predictors by mean absolute SHAP values. Dependence plots were used to explore relationships between feature values and their SHAP attributions. These explanations describe fitted-model behavior, not causal effects, and should be interpreted in light of predictor relationships and the background data [31-32].

3. Results and Discussion

3.1. Model Performance Comparison

The performance of Logistic Regression, XGBoost, AMFormer Full, and AMFormer Ablation was evaluated using the independent test set. Discrimination was assessed using the area under the receiver operating characteristic curve (ROC-AUC) and average precision (AP), while accuracy, precision, recall, specificity, and F1-score described classification performance at the selected decision thresholds. These complementary metrics provide a broader assessment of predictive performance in clinical prediction studies [18], [25]. The results are summarized in Table 2.

Table 2. Performance comparison of the evaluated models on the independent test set.

Model	ROC-AUC	ROC-AUC 95% CI	AP	AP 95% CI	Accu-racy	Preci-sion	Recall	Speci-ficity	F1-score
Logistic Regression	0.7955	[0.7815, 0.8091]	0.6108	[0.5865, 0.6353]	0.7318	0.5261	0.7113	0.7401	0.6048
XGBoost	0.7971	[0.7837, 0.8106]	0.6172	[0.5942, 0.6415]	0.7320	0.5263	0.7142	0.7392	0.6060
AMFormer Full	0.7985	[0.7843, 0.8121]	0.6214	[0.5971, 0.6461]	0.7417	0.5409	0.6935	0.7612	0.6077
AMFormer Ablation	0.7992	[0.7852, 0.8124]	0.6242	[0.6009, 0.6480]	0.7149	0.5040	0.7541	0.6990	0.6042

As shown in Table 2, the four models achieved relatively similar discrimination performance on the independent test set. ROC-AUC values ranged from 0.7955 to 0.7992, while AP values ranged from 0.6108 to 0.6242. AMFormer Ablation obtained the highest ROC-AUC and AP, followed closely by AMFormer Full. However, the differences between the highest and lowest values were only 0.0037 for ROC-AUC and 0.0134 for AP, indicating relatively small differences in predictive discrimination.

The threshold-dependent classification metrics showed different performance patterns across the models. AMFormer Full achieved the highest accuracy (0.7417), precision (0.5409), specificity (0.7612), and F1-score (0.6077). In contrast, AMFormer Ablation obtained the highest recall (0.7541), indicating that it identified a larger proportion of hypertensive respondents. However, this higher recall was accompanied by lower precision (0.5040) and specificity (0.6990), reflecting a greater proportion of false-positive predictions. These results illustrate the trade-off between identifying positive cases and correctly classifying negative cases, which is particularly relevant when evaluating predictive models for health screening [4], [25].

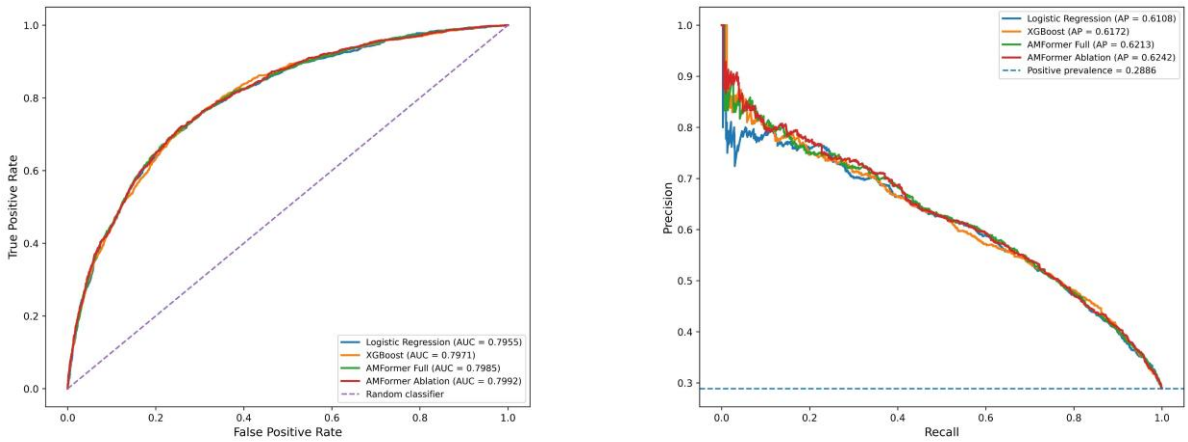


Figure 2. Curves for four models on the independent test set: (a) Receiver operating characteristic (ROC) (b) Precision–recall

The ROC curves in Figure 2(a) show substantial overlap across the four models, visually supporting the small differences in ROC-AUC reported in Table 2. Although AMFormer Ablation achieved the highest value (0.7992), its curve showed no clear separation from the other models. The bootstrap confidence intervals also overlapped across all four models, suggesting considerable uncertainty around the differences in their estimated discrimination performance. Nevertheless, overlapping confidence intervals should not be interpreted as a formal statistical test of equivalence or as proof that the models have identical performance [18], [25].

The precision–recall curves in Figure 2(b) also exhibit similar patterns, particularly across the middle and higher recall ranges. More noticeable fluctuations occur at lower recall values, where precision varies more substantially. The AP values ranged from 0.6108 to 0.6242, with all four models performing above the positive-class prevalence baseline of 0.2886. This indicates that the models achieved better precision–recall performance than the prevalence-based reference. However, performance above this baseline alone does not establish clinical usefulness or readiness for practical screening [25].

Previous studies have evaluated machine-learning approaches for hypertension prediction using Indonesian health data [4], [5] and independent nationwide cohorts in other populations [33-34].

Sismanto et al. [35] also compared ensemble learning with conventional regression for vehicle-inspection forecasting, reporting the strongest performance from Gradient Boosting. Other studies have applied machine-learning classification and logistic regression to health-related and socioeconomic outcomes [7], [36]. Although these studies illustrate the broad application of predictive modelling, differences in populations, outcomes, and predictors limit direct numerical comparisons.

Overall, the results do not demonstrate a substantial predictive advantage of AMFormer over Logistic Regression or XGBoost in this experiment. Although AMFormer Ablation achieved the highest ROC-AUC and AP, AMFormer Full obtained the highest accuracy and F1-score. The small differences in discrimination and overlapping confidence intervals suggest that the evaluated models performed comparably on this test set. Conventional models may therefore remain practical alternatives when computational simplicity is important; however, computational cost and clinical utility were not directly measured in this study. Further external validation and clinical evaluation are needed before any model can be considered suitable for practical hypertension screening.

3.2 Cross-Validation Results and Ablation Analysis

Repeated stratified three-fold cross-validation with three repetitions was conducted using the training set, resulting in nine validation splits. This procedure was used to examine the consistency of AMFormer performance across different data partitions and to assess the contribution of the multiplicative interaction component. For each split, the model checkpoint was selected based on validation ROC-AUC, while the independent test set remained excluded from cross-validation and model selection.

Table 3. Repeated stratified cross-validation performance of AMFormer variants

Model	ROC-AUC (mean ± SD)	Average Precision (mean ± SD)
AMFormer Full	0.8016 ± 0.0038	0.6249 ± 0.0100
AMFormer Ablation	0.8016 ± 0.0039	0.6260 ± 0.0094
Wilcoxon p-value	0.820312	0.164062

As presented in Table 3, the two AMFormer variants achieved nearly identical mean ROC-AUC values. AMFormer Full obtained a mean ROC-AUC of 0.801574 (SD = 0.0038), while AMFormer Ablation obtained 0.801582 (SD = 0.0039). The mean difference was only −0.000008, indicating an almost negligible difference in discrimination performance across the validation splits.

A similar pattern was observed for average precision. AMFormer Full achieved a mean AP of 0.624917 (SD = 0.0100), whereas AMFormer Ablation obtained a slightly higher value of 0.626030 (SD = 0.0094). The difference of −0.001113 in favor of the ablated model was small. The standard deviations were also comparable, suggesting similar observed variability in performance across the nine validation splits.

The ablation experiment was designed to examine whether the multiplicative interaction mechanism provides additional predictive value within the AMFormer architecture. The full configuration included the multiplicative MemoryBlock, whereas the ablated configuration removed this component while retaining standard attention and feed-forward processing. This comparison was intended to isolate the contribution of the multiplicative interaction mechanism under otherwise consistent modelling conditions [11].

The results did not indicate a consistent performance advantage from retaining the multiplicative MemoryBlock. Although the mean ROC-AUC values were virtually identical, AMFormer Ablation achieved a marginally higher mean AP. These findings suggest that, under the present experimental configuration, the multiplicative interaction component did not provide a substantial improvement in

predictive performance. However, this observation is specific to the dataset, predictor set, and model configuration examined in this study and should not be generalized to other tabular learning tasks.

Exploratory paired Wilcoxon signed-rank tests produced p-values of 0.820312 for ROC-AUC and 0.164062 for AP. Neither comparison indicated a statistically significant difference at the 0.05 level. Nevertheless, these p-values should be interpreted cautiously because repeated cross-validation produces correlated performance estimates when training subsets overlap across validation splits. Such dependence can affect the reliability of conventional statistical comparisons and may lead to misleading conclusions about model superiority [28], [37-38].

Overall, AMFormer Full and AMFormer Ablation demonstrated highly comparable cross-validation performance. The negligible difference in mean ROC-AUC and the small difference in AP provide no clear evidence that the multiplicative MemoryBlock improved predictive performance in this experiment. The findings therefore indicate that the additional architectural component did not yield a substantial benefit under the evaluated conditions, while further studies across different datasets and model configurations would be needed to establish its broader contribution.

3.3. SHAP-Based Model Interpretation

SHAP analysis was conducted to interpret AMFormer Full predictions using 1,000 observations from the independent test set and 200 training observations as the background dataset. The `PermutationExplainer` used 14 predictors and a maximum of 29 evaluations per observation. Global importance was summarized using mean absolute SHAP values, while summary and dependence plots were used to examine feature contributions and their patterns [29–32].

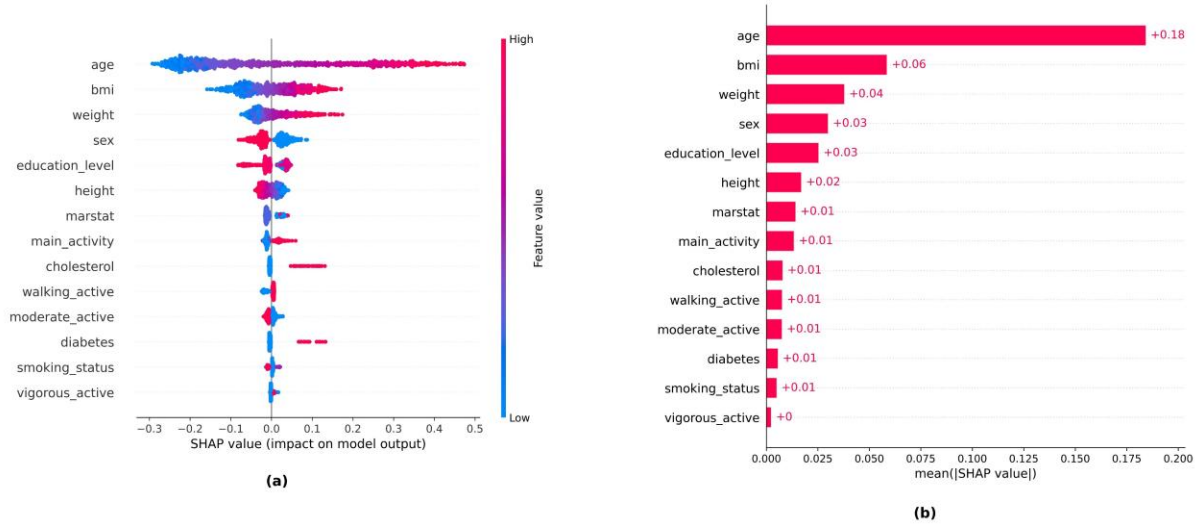


Figure 3. SHAP-based interpretation of the AMFormer Full model on the test set: (a) SHAP summary plot and (b) global feature importance based on mean absolute SHAP values.

As shown in Figure 3(a), age exhibited the widest distribution of SHAP values among the 14 predictors. BMI and weight also showed relatively broad distributions, followed by sex and education level. The remaining predictors generally had SHAP values concentrated closer to zero. These patterns indicate that the AMFormer Full model relied more strongly on age and selected anthropometric and sociodemographic characteristics when generating predictions in the analyzed test sample.

Figure 3(b) presents the global feature importance ranking based on mean absolute SHAP values obtained using the 200-observation background dataset. Age ranked highest (0.186108), followed by BMI (0.058974), weight (0.037398), sex (0.031215), and education level (0.025321). Height ranked sixth (0.016328), followed by marital status (0.014119) and main activity (0.013375). The remaining predictors had smaller mean absolute SHAP values, with vigorous physical activity showing the

lowest importance (0.002429). This ranking reflects the average magnitude of each feature's contribution to the model output across the explanation sample, regardless of contribution direction [30–32], [39].

The SHAP summary plot also illustrates the direction of feature contributions. Higher values of age, BMI, and weight were generally associated with positive SHAP values, whereas lower values more frequently contributed negatively. Height showed an opposing pattern, with higher values generally associated with more negative contributions. These patterns describe the fitted model's predictions rather than independent relationships with hypertension status.

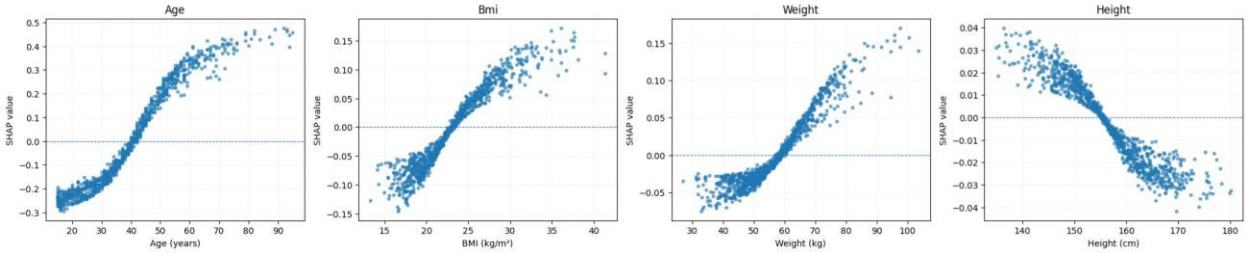


Figure 4. SHAP dependence plots for age, BMI, weight, and height

Figure 4 presents the dependence plots for the four numerical predictors selected based on their global importance. Age showed increasing SHAP values, shifting from predominantly negative contributions at younger ages to positive contributions at older ages. BMI and weight also showed increasing patterns, whereas height exhibited a decreasing pattern, with higher values generally associated with more negative contributions.

These patterns should be interpreted in the context of relationships among predictors. Since BMI, weight, and height are anthropometrically related, their SHAP contributions may reflect overlapping information rather than independent effects. A positive SHAP value indicates a contribution toward a higher model output relative to the reference output, not a direct estimate of an individual's change in hypertension risk [30-32].

The consistency of global feature rankings was assessed using two analyses. First, importance estimates from 300 and 1,000 explanation observations yielded a Spearman rank correlation of 0.9780, with the same five highest-ranked features. This comparison used the initial SHAP computation, in which the background masker automatically subsampled 200 available observations to an effective sample of 100. Second, comparing effective background samples of 100 and 200 observations yielded a correlation of 0.9560. The final plots in Figures 3 used the explicitly configured 200-observation background dataset. These results indicate ranking consistency under the evaluated configurations, but do not establish invariance across all possible sampling strategies. Similar stability considerations have been examined in previous SHAP-based studies [40].

SHAP-based interpretation has also been applied to classification studies using multidimensional survey data. For example, Syukri et al. [10] used SHAP to identify influential predictors in household food resilience classification, demonstrating its application to interpreting gradient-boosting models in a survey-based setting. Although the research context differs, this study illustrates the use of feature attribution to interpret machine-learning predictions from survey data.

Overall, age was the most influential predictor in the AMFormer Full model, followed by BMI, weight, sex, and education level. The stability analyses showed consistent leading feature rankings across the evaluated sample configurations. These findings explain how the model used the available predictors, rather than establishing causal relationships or direct measures of clinical importance.

4. Conclusion

This study compared Logistic Regression, XGBoost, AMFormer Full, and AMFormer Ablation for hypertension classification using 32,413 respondents from IFLS5. On the independent test set, AMFormer Ablation achieved the highest ROC-AUC (0.7992) and average precision (0.6242), while AMFormer Full obtained the highest accuracy (0.7417) and F1-score (0.6077). All models achieved AP values above the positive-class prevalence baseline of 0.2886, although their overall performance differences were small. Repeated cross-validation further showed nearly identical performance between AMFormer Full and AMFormer Ablation, providing no clear evidence of a substantial contribution from the multiplicative MemoryBlock under the evaluated conditions.

SHAP analysis identified age, BMI, weight, sex, and education level as the most influential predictors in AMFormer Full. The stability analyses indicated consistent feature rankings across the evaluated explanation and background sample sizes; however, these attributions describe model behavior rather than causal relationships. Overall, the findings suggest that increased architectural complexity does not necessarily improve hypertension classification when using the available predictors. Further evaluation using more recent data, additional clinical risk factors, and independent population samples is needed to assess model generalizability and potential practical applications.

References

[1] World Health Organization. (2023). Global report on hypertension: The race against a silent killer. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240081062

[2] Chowdhury, M. Z. I., Naeem, I., Quan, H., Leung, A. A., Sikdar, K. C., O'Beirne, M., & Turin, T. C. (2022). Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis. PLOS ONE, 17(4), e0266334.

[3] Silva, G. F. S., Fagundes, T. P., Teixeira, B. C., & Chiavegatto Filho, A. D. P. (2022). Machine learning for hypertension prediction: A systematic review. Current Hypertension Reports, 24(11), 523–533.

[4] Estiko, R. I., Widyantoro, B., Juzar, D. A., Yusup, R. M., Rakhmat, I. F., & Rijanto, E. (2025). Screening hypertension using non-laboratory risk factors with machine learning: A retrospective cross-sectional study in Indonesia. BMJ Open, 15(8), e092364.

[5] Septian, E., Khaefi, M. R., Athoillah, A., Aisyah, D. N., Hardhantyo, M., Rahman, F. M., & Manikam, L. (2025). Prediction of personalised hypertension using machine learning in Indonesian population. Journal of Medical Systems, 49, 137

[6] Alfaqeeh, M., Alfian, S. D., & Abdulah, R. (2023). Factors associated with hypertension among adults: A cross-sectional analysis of the Indonesian Family Life Survey. Vascular Health and Risk Management, 19, 827–836.

[7] Hartono, A., Dewi, L. A., Yuniarti, E., Putri, S. T. H., & Harahap, T. S. (2023). Machine learning classification for detecting heart disease with K-NN algorithm, decision tree and random forest. Eksakta: Berkala Ilmiah Bidang MIPA, 24(4), 513–522.

[8] Ratnaningayu, N. D., Tedjo, A., & Panigoro, S. S. (2024). The implementation of machine learning algorithms for breast cancer biomarker validation in metabolomics studies. Eksakta: Berkala Ilmiah Bidang MIPA, 25(4), 468–483.

[9] Borisov, V., Leemann, T., Sebler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems, 35(6), 7499–7519.

[10] Syukri, N., Oktarina, S. D., & Rahardiantoro, S. (2026). Comparative performance of gradient boosting algorithms for household food resilience classification during the COVID-19. Eksakta: Berkala Ilmiah Bidang MIPA, 27(4), 479–494.

[11] Cheng, Y., Hu, R., Ying, H., Shi, X., Wu, J., & Lin, W. (2024). Arithmetic feature interaction is necessary for deep tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 38(10), 11516–11524.
[12] Zhang, M., Xia, X., Wang, Q., Pan, Y., Zhang, G., & Wang, Z. (2025). Application of machine learning algorithms in predicting new onset hypertension: A study based on the China Health and Nutrition Survey. Environmental Health and Preventive Medicine, 30, 3.
[13] Griselda, M., Alfian, S. D., Wicaksono, I. A., Wawruch, M., & Abdulah, R. (2023). Findings from the Indonesian family life survey on patterns and factors associated with multimorbidity. *Scientific Reports, 13*, 18607.
[14] Strauss, J., Witoelar, F., & Sikoki, B. (2016). The fifth wave of the Indonesia Family Life Survey: Overview and field report (RAND Working Paper No. WR-1143/1-NIA/NICHD). RAND Corporation.
[15] Singh, V., Pencina, M., Einstein, A. J., Liang, J. X., Berman, D. S., & Slomka, P. (2021). Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports, 11*, 14490.
[16] Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Jr., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. BMJ, 384, e074819.
[17] Bey, R., Goussault, R., Grolleau, F., Benchoufi, M., & Porcher, R. (2020). Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. *Journal of the American Medical Informatics Association, 27*(8), 1244–1251.
[18] Allgaier, J., & Pryss, R. (2024). Practical approaches in evaluating validation and biases of machine learning applied to mobile health studies. Communications Medicine, 4, 76.
[19] Efthimiou, O., Seo, M., Chalkou, K., Debray, T. P. A., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ, 386*, e078276.
[20] Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. Anesthesia & Analgesia, 132(2), 365–366.
[21] Harris, J. K. (2021). Primer on binary logistic regression. Family Medicine and Community Health, 9(Suppl 1), e001290.
[22] Ehrman, R. R., & Kline, J. A. (2022). Primer on logistic regression for emergency care researchers. The Journal of Emergency Medicine, 63(5), 683–691.
[23] Wu, Q., Qi, Z., Li, Q., Hao, C., & Tang, S. (2026). Systematic evaluation of machine learning models for clinical risk prediction on real-world hospital datasets. iScience, 29(5), 115654.
[24] Badaro, G., Saeed, M., & Papotti, P. (2023). Transformers for tabular data representation: A survey of models and applications. Transactions of the Association for Computational Linguistics, 11, 227–249.
[25] Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G. M., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., van Smeden, M., Steyerberg, E. W., & Topic Group 6 of the STRATOS Initiative. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. The Lancet Digital Health, 7(12), 100916.
[26] Binuya, M. A. E., Engelhardt, E. G., Schats, W., Schmidt, M. K., & Steyerberg, E. W. (2022). Methodological guidance for the evaluation and updating of clinical prediction models: A systematic review. BMC Medical Research Methodology, 22, 316.
[27] Coley, R. Y., Liao, Q., Simon, N., & Shortreed, S. M. (2023). Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: A case study in suicide risk prediction. BMC Medical Research Methodology, 23, 33.

[28] Sziklai, B. R., Baranyi, M., & Héberger, K. (2025). Does cross-validation work in telling rankings apart? Central European Journal of Operations Research, 33, 1503–1528.
[29] Chen, H., Lundberg, S. M., & Lee, S.-I. (2022). Explaining a series of models by propagating Shapley values. Nature Communications, 13, 4512.
[30] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clinical and Translational Science, 17(11), e70056.
[31] Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: From cooperative game to explainable artificial intelligence. Autonomous Intelligent Systems, 4, 2.
[32] Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), 2400304.
[33] Hwang, S. H., Lee, H., Lee, J. H., Lee, M., Koyanagi, A., Smith, L., Rhee, S. Y., Yon, D. K., & Lee, J. (2024). Machine learning-based prediction for incident hypertension based on regular health checkup data: Derivation and validation in 2 independent nationwide cohorts in South Korea and Japan. *Journal of Medical Internet Research, 26*, e52794.
[34] Ghiasi, S., Darroudi, S., Moradi, M., Kamrani, F., Fatehi, M. R., Sajjadi, S. M., Yousefian, S., Amiri, S., Ghayour-Mobarhan, M., Mahdavizadeh, V., Nasiri, J. A., Mansoori, A., & Esmaily, H. (2026). A comparison of machine learning algorithms for predicting hypertension incidence based on cohort study. *Endocrinology, Diabetes & Metabolism, 9*(3), e70199.
[35] Sismanto, B., Nurhayati, N., Tjahyaningtijas, R. R. H. P. A., Mahmud, J., Figueredo, R. E., & Varshney, A. (2026). Vehicle inspection forecasting through temporal signal processing and ensemble machine learning. *Eksakta: Berkala Ilmiah Bidang MIPA, 27*(3), 431–448.
[36] Nafis, M. A., Susilaningrum, D., Ulama, B., & Kusrini, D. E. (2024). Assessing the impact of household socioeconomic factors on clean and healthy living behaviors with binary logistic regression: A study in Probolinggo Regency. Eksakta: Berkala Ilmiah Bidang MIPA, 25(4), 436–447.
[37] Jafrasteh, B., Adeli, E., Pohl, K. M., Kuceyeski, A., Sabuncu, M. R., & Zhao, Q. (2025). Statistical variability in comparing accuracy of neuroimaging based classification models via cross validation. *Scientific Reports, 15*, 28745.
[38] Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association, 119*(546), 1434–1445.
[39] Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data, 11*, 44.
[40] Tiukhova, E., Vemuri, P., López Flores, N., Islind, A. S., Óskarsdóttir, M., Poelmans, S., Baesens, B., & Snoeck, M. (2024). Explainable learning analytics: Assessing the stability of student success prediction models by means of explainable AI. *Decision Support Systems, 182*, 114229.