Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

APPLICATION OF XGBOOST AND CATBOOST ALGORITHMS FOR ELDERLY HYPERTENSION CLASSIFICATION ON IFLS 5 DATA

Ekklesiafilifi Loyalita Crossesa[1,*)], A'yunin Sofro[2,*)]

[1)]*Department of Mathematics, State University of Surabaya, Surabaya, Indonesian*
[2)]*Actuarial Science Department, State University of Surabaya, Surabaya, Indonesian*
[*]*email: ayuninsofro@unesa.ac.id*

Abstract: *Hypertension in the elderly poses complex classification challenges, characterized by noisy categorical features in health survey datasets. This study focuses on using XGBoost and CatBoost algorithms to overcome barriers when classifying hypertension in the elderly (≥ 60 years) using IFLS 5 data. Unlike standard methods that focus on accuracy, this evaluation emphasizes the recall metric to reduce false negative errors, which is crucial for ensuring safety in medical screening. After carefully tuning the hyperparameters using GridSearchCV and 5-fold cross-validation on 2,774 participants, the models revealed clear algorithmic trade-offs. CatBoost demonstrated superior generalization stability and achieved the highest accuracy (66.49%), while XGBoost exhibited significant superiority in sensitivity (recall of 80.18%) by effectively applying regularization to detect minority class signals. Evaluating feature significance using the information gain and prediction values change metrics verified that biological indicators, particularly diabetes and BMI, were the main predictors compared to demographic variables. In summary, CatBoost is reliable, but XGBoost is better suited for building clinical decision support systems where the priority is detecting sensitivity.*

Keywords: *CatBoost; Classification; Elderly Hypertension; IFLS 5; XGBoost*

Abstrak: Hipertensi pada lansia menimbulkan tantangan klasifikasi yang kompleks, ditandai dengan fitur kategorikal yang berisik dalam dataset survei kesehatan. Penelitian ini berfokus pada penggunaan algoritma *XGBoost* dan *CatBoost* untuk mengatasi hambatan dalam mengklasifikasikan hipertensi pada lansia (≥ 60 tahun) menggunakan data IFLS 5. Berbeda dengan metode standar yang berfokus pada akurasi, evaluasi ini menekankan metrik *Recall* untuk mengurangi kesalahan *False Negative*, yang sangat penting untuk memastikan keamanan dalam skrining medis. Melalui penyesuaian *hyperparameter* yang teliti menggunakan *GridSearchCV* dan validasi silang 5-*fold* pada 2.774 peserta, model-model tersebut menunjukkan kompromi algoritmik yang jelas. *CatBoost* menonjol dalam stabilitas generalisasi dengan akurasi tertinggi (66,49%), sementara *XGBoost* menunjukkan keunggulan yang signifikan dalam sensitivitas (*Recall* 80,18%) dengan menerapkan regularisasi secara terampil untuk mendeteksi sinyal kelas minoritas. Evaluasi signifikansi fitur menggunakan metrik *Information Gain* dan *PredictionValuesChange* memverifikasi bahwa indikator biologis, terutama diabetes dan BMI, merupakan prediktor utama dibandingkan dengan variabel demografis. Secara ringkas, meskipun *CatBoost* menawarkan keandalan, *XGBoost* lebih cocok untuk membangun sistem dukungan keputusan klinis di mana prioritas deteksi sensitivitas sangat penting.


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

Kata Kunci: CatBoost; Hipertensi Lansia; IFLS 5; Klasifikasi; XGBoost

INTRODUCTION

Global demographic changes have made the elderly the fastest growing segment of the population, including in Indonesia. This aging population directly contributes to an increase in chronic diseases, especially hypertension. According to the Nasional Riset Kesehatan Dasar (Riskesdas) report, hypertension prevalence among individuals aged 60 years and older is significantly high, exceeding 50% (Kementerian Kesehatan RI, 2019). This alarming trend is further reinforced by an analysis of Indonesian Family Life Survey (IFLS) data identifying older adults as the most vulnerable demographic group to hypertension in Indonesia (Siagian, 2022).

Hypertension is clinically categorized as a "silent killer" because it does not exhibit symptoms in the early stages. Therefore, effective early detection mechanisms are necessary to prevent fatal complications, such as stroke and heart failure (World Health Organization, 2021). However, large-scale early screening that relies on manual medical diagnosis is resource-intensive and prone to human error. Therefore, developing an automatic classification model based on risk factors is mathematically necessary to efficiently and accurately diagnose and detect hypertension, especially in environments with limited resources (Chowdhury et al., 2022).

Significant computational challenges often arise when developing reliable classification models for health survey data, such as the Indonesian Family Life Survey (IFLS). High dimensionality, noise, and the dominance of categorical features representing socioeconomic and lifestyle factors are typical characteristics of medical data. These variables often interact with biological indicators in complex nonlinear ways, making them difficult to model effectively using traditional parametric statistics, such as logistic regression, due to their sensitivity to multicollinearity and strict assumptions about linearity (Uddin et al., 2019; Kurniawan et al., 2023). Therefore, a nonparametric approach that can analyze complex patterns directly from the data without relying on rigid distribution assumptions is needed to address this structural complexity and drive a strategic shift toward machine learning (ML) techniques. A more adaptive approach with Machine Learning Algorithms is strongly supported by a literature review showing that decision tree-based algorithms and their development through Ensemble Learning can produce better predictions on diverse medical data (Sarkar, 2020).

Ensemble Learning algorithms have emerged as the leading solution for structured tabular data. In this field, their ability to optimize classification evaluation results has made Gradient Boosting decision trees highly desirable. Extreme Gradient Boosting (XGBoost) is widely used due to its application of second-order Taylor expansion for objective function approximation and advanced regularization (L_1 and L_2) in preventing overfitting, making it highly efficient for sparse data (Chen & Guestrin, 2016). On the other hand, Categorical Boosting (CatBoost) offers uniqueness in handling categorical features through Ordered Target Statistics and Oblivious Trees as advantages, which are designed to reduce prediction shifts and overfitting often caused by standard target encoding (Prokhorenkova et al., 2018; Hancock &


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

Khoshgoftaar, 2020). The use of these two Ensemble Learning algorithms is crucial for observing how classification results with optimal hyperparameter tuning can achieve better convergence on geriatric survey data.

Although boosting algorithms are popular, existing literature on disease classification often ignores cost-specific sensitivity in medical diagnosis. Previous studies have explored algorithms such as Support Vector Machine and Naive Bayes for classifying hypertension (Lathifah & Pratiwi, 2022). However, these studies focused on maximizing overall accuracy or the area under the curve (AUC) (Handayani et al., 2018; Islam et al., 2023). This approach has created a significant research gap because, in clinical settings, minimizing false negatives (type II errors) is far more critical than maximizing overall accuracy. Models with high accuracy are problematic for medical screening if they fail to detect positive cases because missed hypertensive patients can lead to treatment failure and serious health consequences (Hossin & Sulaiman, 2015). Recent studies emphasize that evaluation metrics must align with clinical objectives. However, the direct comparison of XGBoost and CatBoost optimized for recall (sensitivity) on Indonesian elderly data remains limited (Hicks et al., 2022).

This study aimed to fill this gap by analyzing the application of XGBoost and CatBoost algorithms to classify hypertension in the elderly using IFLS 5 data. The study also aimed to determine which algorithm is superior at grouping hypertension cases in elderly individuals. This study's main contribution lies in its methodological evaluation of how different boosting optimization schemes perform under the constraint of maximizing recall. Specifically, it compares XGBoost's regular objective function with CatBoost's ordered boosting. Additionally, feature importance analysis based on the gain metric for XGBoost and the prediction values change approach for CatBoost was used to interpret the mathematical weights of biological and demographic predictors. Prioritizing sensitivity (recall) over accuracy provides a computational perspective on developing safer, more effective, and systematic clinical decision support systems.

RESEARCH METHODS

The data used in this study were obtained from the Indonesian Family Life Survey (IFLS) Wave 5. This large-scale, longitudinal survey examined the socioeconomic and health dynamics of a population representing 83% of Indonesia's population. It was conducted from 2014 to 2015 (Strauss et al., 2016). The research sample focused on elderly respondents ($\geq$ 60 years) and the selected variables were based strictly on Joint National Committee 8 (JNC 8) clinical guidelines and the latest population studies on geriatric hypertension. The dependent variable (Y) is hypertension status (0: normal; 1: hypertension), which was determined based on systolic blood pressure $\geq$ 140 mmHg and diastolic blood pressure $\geq$ 90 mmHg (James et al., 2014). The independent variable (X) consists of ten predictors designed to comprehensively cover hypertension risk factors, especially among the elderly. These factors include demographic aspects (age, gender, and education); biometric factors (BMI); lifestyle factors (smoking and sleep quality); and clinical history factors (diabetes, cholesterol, and heart attack).


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

These variables were selected based on the availability of IFLS 5 longitudinal data, their alignment with medical guidelines, and their ability to capture the multifactorial nature of hypertension and address complexities often overlooked in simpler models (Kurniawan et al., 2023).

Table 1. Research Variables Description

Variables		Information	Type
Y	Hypertension	Target Variable (0: Normal, 1: Hypertension)	Categorical
X_1	BMI	Body Mass Index	Numeric
X_2	Age	Individual age (Year)	Numeric
X_3	Sex	Individual gender (0: Female, 1: Male)	Categorical
X_4	Employment_Status	Employment status (0: Not working, 1: Working)	Categorical
X_5	Education_Level	Highest level of education (0: No schooling, 1: Primary school equivalent, 2: Junior high school equivalent, 3: Senior high school equivalent, 4: Higher education)	Categorical
X_6	Smoking_Status	Smoking Status (0: No, 1: Smoker)	Categorical
X_7	Sleep_Quality	Sleep Quality (0: Insufficient, 1: Adequate)	Categorical
X_8	Heart_Attack	History of Cardiovascular Comorbidities (0: Normal, 1: HeartAttack)	Categorical
X_9	Diabetes	History of Metabolic Comorbidities (0: Normal, 1: Diabetes)	Categorical
X_{10}	High_Cholesterol	History of Metabolic Comorbidities (0: Normal, 1: HighCholesterol)	Categorical

Data Source: RAND IFLS Wave 5

The stages of data analysis in this study were systematically carried out in four main stages: (1) data preprocessing, including cleaning the data and handling missing values to maintain the statistical distribution, (2) splitting the data for training and testing with an 80:20 ratio to reduce bias and evaluate generalization, (3) training the model with hyperparameter optimization and cross-validation, and (4) interpreting the results of applying classification algorithms (XGBoost and CatBoost) and feature importance extraction. The 80:20 data split was chosen based on the Pareto Principle and empirical evidence showing that allocating 80% of the data for training provides algorithms with sufficient variance to learn complex non-linear patterns, while the remaining 20% ensures a statistically significant sample size for confusion


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

matrix validation, thereby minimizing the variance in performance estimates (Gholamy et al., 2018; Nguyen et al., 2021).

The classification model in this study uses two gradient boosting frameworks. The first is Extreme Gradient Boosting (XGBoost). XGBoost is an extension of Gradient Boosting Machine (GBM) that optimizes the objective function through second-order Taylor expansion based on the principle of additive learning mathematically, which provides faster convergence compared to traditional GBM that only uses first-order derivatives (Friedman, 2001). If $\hat{y}_i^{(t)}$ represents the prediction at iteration t, then the objective function to be maximized is:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (1)$$

Where g_i and h_i are the first and second gradients of the loss function. The theoretical advantage of XGBoost lies in the regularisation component $\Omega(f_t)$, which controls the complexity of the model to prevent overfitting.

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{1}{2} \lambda \|w\|^2 \quad (2)$$

Where T is the number of leaves and w is the leaf weight. This L_1 and L_2 regularization makes XGBoost highly robust against noise in medical data as a prevention, such as IFLS 5 data (Chen & Guestrin, 2016). In the context of degenerative disease detection, this ability is crucial to minimize false negatives, making it safer to use as a screening tool and making the XGBoost algorithm a viable choice in this study.

The second algorithm is Categorical Boosting (CatBoost). CatBoost addresses the target leakage problem in conventional boosting algorithms, which are commonly found in survey data, by using the Ordered Target Statistics method. Rather than using one-hot encoding, which simple algorithms often use to increase data dimension (sparsity) and which often causes target leakage (prediction bias), CatBoost converts the x_k category into a numerical value based on the expected target y.

$$\hat{x}_i^k = \frac{\sum_{j=1}^{p-1} \left[x_j^k + x_i^k \right] \cdot y_j + a \cdot P}{\sum_{j=1}^{p-1} \left[x_j^k + x_i^k \right] + a} \quad (3)$$

Here, P is the prior value, which is usually the target mean in the dataset, and a is the prior weight. This equation ensures that the encoding of the current sample depends solely on the observed history, thereby eliminating the bias that occurs when data is used to calculate statistics (Prokhorenkova et al., 2018). This makes CatBoost highly efficient for datasets such as IFLS that are dominated by categorical variables or questionnaires (Hancock & Khoshgoftaar, 2020).

To optimise model performance, hyperparameter tuning was used with GridSearchCV, a method that works by defining a grid of hyperparameter values to be tested, then training the model for each combination using cross-validation (CV) (Pedregosa et al., 2011). Model


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

validation is performed repeatedly to assess model generalisation, avoid overfitting on the data, and thus reduce variance bias ($k = 5$) (Hastie et al., 2009).

A feature importance analysis was performed after model training to interpret the logic behind the classification decisions. Different metric definitions were used in this study, along with optimization scheme adjustments in each algorithm, to strictly measure the contribution of each predictor.

1. For the XGBoost algorithm, feature importance was evaluated using the Gain metric, which represents the average increase in accuracy or decrease in objective function generated by a particular feature across the decision tree. Features with the highest Gain values are considered most significant in distinguishing between hypertensive and non-hypertensive classes (Chen & Guestrin, 2016). Mathematically, the difference in structural values after separation with a penalty in the form of a regularization term γ, is the gain calculation for a specific separating node j, using feature k.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \quad (4)$$

   It can be seen that G and H represent the sum of the first- and second-order derivatives (gradient and Hessian, respectively) of the loss function for the left (L) and right (R) child nodes, and that the L_2 regularization parameter is represented by the symbol λ. Features that are more important in minimizing classification error are indicated by a higher total increase.
2. Meanwhile, in CatBoost, feature importance is determined using the PredictionValuesChange approach, which assesses the average change in model prediction values when feature values are altered. This approach has proven to be very useful for evaluating the contribution of categorical features in a boosted tree framework (Hancock & Khoshgoftaar, 2020). Unlike metrics, which are based solely on the frequency of splits, this approach considers the magnitude of the impact on leaf values. This impact is calculated by summing the weighted variance of the leaf values (v) for each split involving that feature.

$$Importance_F = \sum_{trees,leaves} \left(c_1\left(v_1 - v_{avg}\right)^2 + c_2\left(v_2 - v_{avg}\right)^2\right) \quad (5)$$

   c_1 and c_2 represent the instance weights in the left and right leaves, respectively, and v_1 and v_2 represent the leaf values. v_{avg} represents the weighted average of the leaf values. This method provides a powerful measure of feature influence and is particularly effective for models dealing with categorical data.


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

Evaluation metrics used in classification typically include accuracy, sensitivity or recall, specificity, precision, and F1 score. The main evaluation metric used in this study is Recall (Sensitivity), which is defined as the proportion of correct positive predictions and minimizing false negatives is critical in medical screening applications (Hicks et al., 2022).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{6}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \times 100\% \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)TP + TN}{Precision + Recall} \times 100\% \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{9}$$

To ensure that the performance differences between XGBoost and CatBoost were statistically significant and not due to random data partitioning, 95% Confidence Interval (CI) were calculated for each metric. Based on the variance observed at $k = 5$ cross validation folds, the CI calculation used the following formula:

$$CI = \bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{k}} \tag{10}$$

Where $\bar{x}$ is the average score, s is the standard deviation across folds, and k is the number of folds. While $t_{n-1,\alpha/2}$ (Z) is the critical value for the 95% confidence level ($Z = 1.96$). This statistical validation is important because it assesses model stability and generalization error, providing a robust measure of reliability beyond single-point estimates (Raschka, 2020; Berrar, 2019).

Then, the area under the ROC curve (AUC) was used to evaluate the model's ability to discriminate at various thresholds (Fawcett, 2006).

$$AUC = \frac{Sensitivity + Specificity}{2} \times 100\% \tag{11}$$

Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

RESULT AND DISCUSSION

RESULT

From the large amount and variety of data scattered throughout IFLS 5, relevant data was collected and combined as dependent and independent variables for the study using the primary key column, namely pidlink. The independent variables for this study were obtained from several questionnaire books in IFLS wave 5, such as Book US for all aspects related to anthropometrics; Books 3A and 3B for demographic, socioeconomic, and medical history information.

The accuracy of the classification results is highly dependent on the quality of the input data. Given that the data from the Indonesian Family Life Survey (IFLS 5) is complex and full of disturbances, pre-processing steps are very important. This stage involves cleaning the data to remove extreme values and handling missing data through appropriate statistical imputation techniques, such as mode for categorical variables and median for numeric variables. This screening process yielded a sample of 2,774 elderly individuals for analysis, comprising 1,666 elderly individuals with hypertension and 1,108 elderly individuals with normal status. These results show that the cleaned data has a 60:40 ratio for the target variable. In the real world, a 60:40 ratio for medical data is within the tolerance threshold for tree-based algorithms, such as XGBoost and CatBoost, without requiring heavy synthetic sampling.

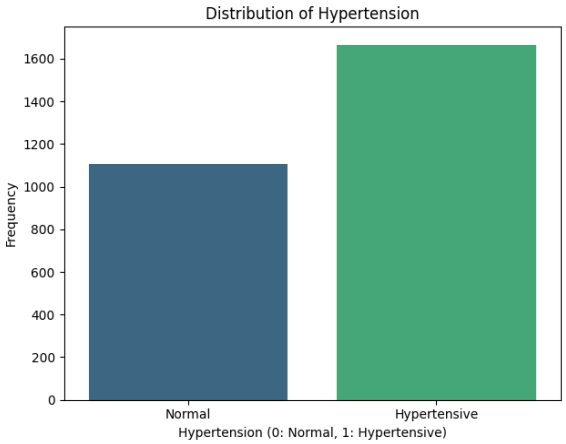

Figure 1. Distribution of Elderly Hypertension Data

During the model training phase, data processing was not performed using standard settings. Rather, it was done through careful hyperparameter tuning using grid search with five-fold cross-validation to find the best parameter configuration. The details of the optimal hyperparameter configuration for hypertension classification in the elderly are provided below.

Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

Table 2. Best Hyperparameter of Machine Learning Approach (GridSearchCV)

Ensemble Boosting Algorithm		Value
XGBoost	colsample_bytree	0.9
	learning_rate	0.01
	max_depth	3
	n_estimators	200
	subsample	0.9
CatBoost	depth	4
	iterations	100
	l2_leaf_reg	5
	learning_rate	0.05

The optimization process produced a configuration with a relatively shallow tree structure, a conservative learning rate, stochastic subsampling, and 200 estimators. This shows that XGBoost requires strong regularization to prevent overfitting on noisy IFLS data. In contrast, CatBoost achieved optimal performance with a slightly deeper structure and higher learning rate over 100 iterations. A substantial L_2 regularization term was selected, confirming the algorithm's internal mechanism for controlling model complexity when handling categorical features.

Table 3. Classification Performance Metrics (Mean $\pm$ 95% CI)

Ensemble Boosting Algorithm	Accuracy	Precision	Recall	F1-Score	ROC - AUC
XGBoost	0.6595 [61.80 – 69.73%]	0.6796 [63.32 – 72.71%]	0.8018 [75.68 – 84.47%]	0.7357 [69.86 – 77.20%]	0.7138
CatBoost	0.6649 [62.34 – 70.45%]	0.6994 [65.04 – 74.50%]	0.7591 [70.93 – 80.49%]	0.7281 [68.94 – 76.54%]	0.7196

Table 3 presents a comparison of model performance in classifying hypertension in the elderly in the optimized IFLS 5 test data. Although CatBoost shows slightly better stability in overall accuracy (0.6649) and Area Under the Curve (AUC: 0.7196), XGBoost shows a clear advantage in the main metric of concern, namely Recall of 0.8018.

Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

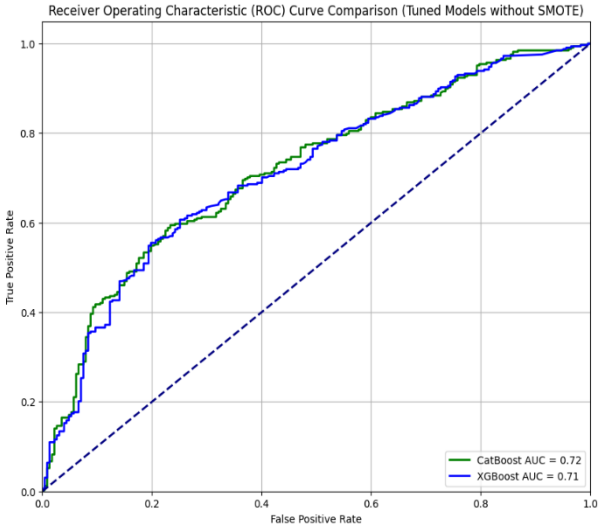

Figure 2. XBoost & CatBoost ROC Curve

The 95% Confidence Interval (CI) is used as a statistical significance assessment of the performance difference between two classification algorithms. As shown in Table 3, XGBoost significantly outperforms CatBoost, whose Recall only reaches 0.7591. Although there is a slight overlap in the confidence interval, the distribution of XGBoost sensitivity clearly shifts towards the upper limit.

This difference of approximately 4.3% in Recall is clinically and practically significant. In the context of screening a large elderly population, a 4.3% increase in sensitivity (recall) means that for every 1,000 positive hypertension cases, XGBoost successfully identifies approximately 43 more patients than CatBoost. These are patients who should have been classified as "healthy" (False Negatives) and missed the opportunity for early intervention. Therefore, XGBoost provides a safer and more effective solution for medical screening purposes where minimizing false negatives is crucial, even though CatBoost has a marginal advantage in overall accuracy.

To support the classification results, feature importance analysis was performed to understand the medical logic behind the model predictions, calculated using the Gain metric for XGBoost and PredictionValuesChange for CatBoost. The results are visualized in Figure 3, which shows strong consensus between the two algorithms regarding the most influential predictors.

Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

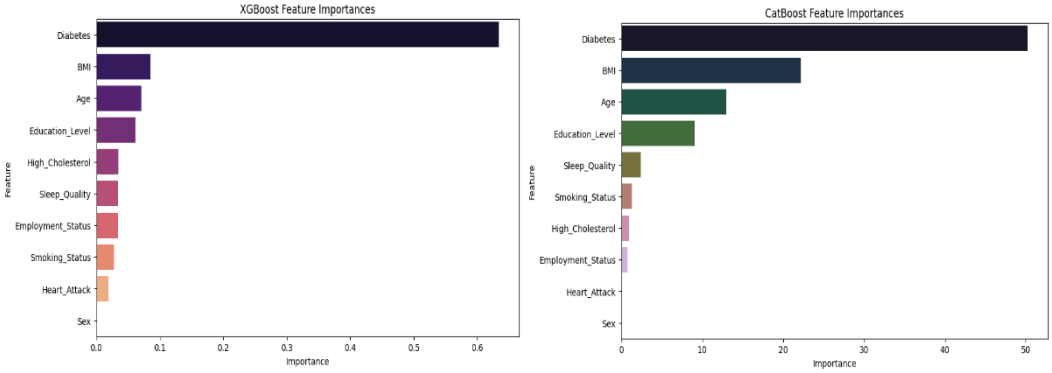

Figure 3. XBoost & CatBoost Features Importance Graph

The findings show consistency in both algorithms that biological risk factors dominate the classification hierarchy. 'Diabetes' emerged as the single most important predictor in both models (XGBoost ≈ 0.63; CatBoost ≈ 50.21), followed by 'BMI' (XGBoost ≈ 0.086; CatBoost ≈ 22.23) as the second-ranked predictor, despite a significant difference. Interestingly, demographic variables such as 'Gender' showed zero contribution in both algorithms, meaning that gender is not a discriminating factor for hypertension risk among the Indonesian elderly population in this dataset.

DISCUSSION

The findings if this research highlights a clear trade-off in computation between CatBoost's generalization stability and XGBoost's sensitivity. Although CatBoost achieves slightly better accuracy and AUC, XGBoost excels in recall (80.18%), a key metric in this study aimed at reducing false negatives. This advantage stems from XGBoost's objective function, which includes L_1 (lasso) and L_2 (ridge) regularization terms (Ω(f_t)). These terms penalize model complexity, preventing decision trees from overfitting the majority class (normal) and enabling the algorithm to detect finer patterns in the minority class (hypertension) (Ogunleye & Wang, 2020). Conversely, while the Ordered Boosting CatBoost approach is effective in minimizing prediction shifts and smoothing decision boundaries for overall accuracy (Hancock & Khoshgoftaar, 2020), it proved too cautious for this dataset. This resulted in more missed positive cases compared to XGBoost.

Hyperparameter tuning provides additional details about the level of model complexity required for the IFLS 5 dataset. The optimal XGBoost settings are shallow trees (max_depth = 3) combined with a moderate learning rate of 0.01. These settings suggest that the data structure is highly nonlinear but noisy. Shallow trees act as "weak learners" with significant bias and minimal variance; however, when gradually improved with a slow learning rate, they can correct errors without capturing the random noise commonly found in survey data (Sagi & Rokach, 2018). This contradicts the assumption that deeper trees are essential for complex medical data. Our observations align with those of Zhang et al. (2020), who recommend prioritizing


Leibniz: Jurnal Matematika
Januari 2026
Vol. 6 No. 01 Hal. 01-14
e-ISSN: 2775-2356
https://ejurnal.unisap.ac.id/leibniz/index

controlling tree depth over expanding ensemble size for tabular health data to ensure model reliability.

From a statistical perspective, the approximately 4.3% difference in recall is supported by the non-overlapping distribution at the upper end of the detailed 95% confidence interval in the results. In practical terms, this difference is significant. Unlike previous hypertension prediction studies, such as that of Handayani et al. (2018), which emphasized overall accuracy of up to 78%, our study argues that a model with 66% accuracy and 80% recall has greater clinical value for screening purposes. High-accuracy models that miss positive cases create a dangerous illusion of safety for patients. By prioritizing recall optimization, XGBoost identifies more at-risk elderly individuals who require treatment, thus reinforcing its role as an effective initial screening method (Hicks et al., 2022).

In terms of feature evaluation, the strong influence of "Diabetes" and "BMI" on the "Information Gain" (XGBoost) and "Prediction Values Change" (CatBoost) measures indicates that continuous variables with significant variance play the largest role in reducing entropy at the split point. Notably, "Gender" (sex) has no impact on either model. This finding challenges some traditional medical perspectives but aligns with computational logic. Categorical features with limited options generally yield lower information gain than continuous variables in gradient boosting systems unless the target is significantly impacted by that category (Lundberg et al., 2020). Therefore, for Indonesian older adults in the IFLS 5 sample, biological indicators (e.g., metabolic health) are much stronger predictors than demographic factors.

CONCLUSION

This study concludes that, although CatBoost offers the highest overall accuracy and better generalization stability, XGBoost is the best computational option for screening hypertension in elderly individuals due to its significantly higher sensitivity (recall: 80.18%). By using a regularized objective function to reduce false negatives, XGBoost provides a more reliable clinical decision support system than CatBoost, which is more cautious. Furthermore, feature analysis shows that physiological indicators, particularly diabetes and BMI, play a major role in influencing hypertension risk, far exceeding demographic elements. However, this study is limited by its use of cross-sectional IFLS 5 data, which provides only a snapshot of health conditions at a single point in time, hindering the modeling of blood pressure changes over time. Additionally, the omission of genetic biomarkers limits the biological completeness of the prediction model. Future studies should prioritize applying Deep Learning models to longitudinal datasets to track disease progression over time. Additionally, using hybrid hyperparameter tuning methods, such as Bayesian optimization or evolutionary algorithms, is recommended to surpass the existing grid search standard and improve algorithm efficiency.

REFERENCES

Berrar, D. (2019). Performance measures for machine learning classification. In Encyclopedia of bioinformatics and computational biology (pp. 988–994). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chowdhury, M. Z. I., et al. (2022). Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis. PLOS ONE, 17(4), e0266334. https://doi.org/10.1371/journal.pone.0266334

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation (Technical Report UTEP-CS-18-09). El Paso, TX: University of Texas at El Paso.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. Journal of Big Data, 7(1), 94. https://doi.org/10.1186/s40537-020-00369-8

Handayani, A., et al. (2018). Hypertension prediction system using data mining techniques. Journal of Physics: Conference Series, 1196.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer.

Hicks, S. A., et al. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979. https://doi.org/10.1038/s41598-022-09966-1

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.

Islam, M. M., et al. (2023). Predicting the risk of hypertension using machine learning algorithms: A cross-sectional study in Ethiopia. PLOS ONE, 18(8), e0289613.

James, P. A., et al. (2014). 2014 evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA, 311(5), 507-520.

Kementerian Kesehatan RI. (2019). Laporan nasional Riskesdas 2018. Jakarta: Badan Penelitian dan Pengembangan Kesehatan.

Kurniawan, R., et al. (2023). Hypertension prediction using machine learning algorithm among Indonesian adults. Journal of Big Data.

Lathifah, N. B., & Pratiwi, D. (2022). Komparasi algoritma Support Vector Machine dan Naive Bayes untuk klasifikasi penyakit hipertensi. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 6(1), 127–133.

Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67.

Nguyen, Q. H., et al. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021.

Ogunleye, A., & Wang, Q. G. (2020). XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(6), 2131-2140.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

Raschka, S. (2020). Model evaluation, model selection, and algorithm selection in machine learning. Cognitive Computation, 12(1), 1063–1093. https://doi.org/10.1007/s12559-020-09740-9

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

Sarkar, J. (2020). Practical machine learning with Python: A problem-solver's guide to building real-world intelligent systems. New Delhi: BPB Publications.

Siagian, T. H. (2022). Hipertensi pada lansia di Indonesia: Tinjauan data IFLS 5. Jurnal Epidemiologi Kesehatan Komunitas, 7(1).

Strauss, J., Witoelar, F., & Sikoki, B. (2016). The Fifth Wave of the Indonesia Family Life Survey (IFLS5): Overview and field report. Santa Monica, CA: RAND Corporation.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1), 281.

World Health Organization. (2021). Hypertension. Retrieved from https://www.who.int/news-room/fact-sheets/detail/hypertension

Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., & Lyashevska, O. (2019). Predictive analytics with gradient boosting trees in clinical medicine. Annals of Translational Medicine, 7(7).