IJIE

INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

Vol. 18 No. 4 (2026) 79-92
https://publisher.uthm.edu.my/ojs/index.php/ijie

# Robustness Evaluation of Machine Learning and Deep Learning Models for Patient Adherence Classification under Feature-Homogeneous Clinical Data

**Silfiana Nur Hamida[1], Wayan Firdaus Mahmudy[2]***

[1] *Artificial Intelligence Innovation Center, University of Brawijaya, Malang 65145, INDONESIA*
[2] *Department of Informatics Engineering, Faculty of Computers Science, University of Brawijaya, Malang 65145, INDONESIA*

*Corresponding Author: wayanfm@ub.ac.id
DOI: https://doi.org/10.30880/ijie.2026.18.04.006

**Article Info**

Received: 9 February 2026
Accepted: 30 April 2026
Available online: 17 June 2026

**Keywords**

Machine learning, clinical patient adherence, model robustness, homogeneous, classification performance

**Abstract**

Model performance in clinical classification is strongly influenced by dataset characteristics, particularly feature homogeneity and interaction complexity. This study evaluates Naive Bayes, Support Vector Machine, Random Forest, XGBoost, CatBoost, LightGBM, Extreme Learning Machine, and a tree-based ensemble on patient adherence and health-risk claim datasets. Experimental results show that on the patient adherence dataset, all models exhibit convergent performance, with accuracies ranging from 0.5816 to 0.6232, indicating limited separability and weak nonlinear structure. In contrast, the health-risk claim dataset presents a more complex feature space, where boosting-based models achieve superior performance: CatBoost 0.8333, LightGBM 0.8305, and XGBoost 0.8253, while Naive Bayes significantly degrades to 0.3356. These findings highlight the importance of feature interaction complexity in determining model effectiveness. Robustness is evaluated by measuring relative performance degradation across datasets under distributional shift, comparing performance on the simpler dataset with that on the more complex one. Boosting-based models demonstrate stable robustness between 30.28% and 31.98%, whereas Naive Bayes exhibits severe negative robustness 67.33%. The ensemble model achieves a lower robustness 24.73%, indicating limited benefit from combining highly correlated learners. These results demonstrate that boosting-based models are more resilient to dataset shifts because they can capture nonlinear feature interactions, whereas probabilistic models are highly sensitive to violations of independence assumptions. The main contribution of this study is a cross-dataset robustness evaluation that quantifies model sensitivity under structural variation. The findings further show that feature homogeneity and interaction complexity are more influential than algorithm selection in determining model performance and robustness.

## 1. Introduction

Patient adherence to medication is crucial for treatment effectiveness. Low adherence can lead to health complications and increased healthcare costs. Clinical datasets that capture patient adherence often show highly similar feature distributions across classes, making it hard for models to distinguish meaningful patterns [1]. This challenge is more pronounced in tabular datasets with limited variability. Both machine learning and deep learning models struggle to capture discriminative information in these cases [2] [3]. Modern machine learning algorithms, such as Random Forest, SVM, Naive Bayes, XGBoost, CatBoost, LightGBM, and ensemble methods, address non-linear relationships through adaptive weighting and multi-level splitting. Lightweight deep learning models, including ELMs, use layered feature representations [4].

To address these challenges, this study makes three main contributions. First, it introduces a quantitative framework to evaluate model robustness based on performance changes across clinically related datasets with differing structural characteristics. Second, it provides a systematic analysis of the impact of feature homogeneity on the discriminative capability of models, an aspect that has been relatively underexplored in prior research. Third, it examines the interaction between feature engineering strategies and dataset structure, demonstrating that dataset characteristics play a more dominant role than algorithmic complexity in determining classification robustness.

Previous studies have shown that tree-based and boosting models often outperform deep learning approaches on tabular data [5]. However, these evaluations are still limited to a single dataset and, therefore, have not been able to adequately capture changes in model performance under dataset shift conditions across clinical datasets with different structures [5]. In addition, prior research has indicated that machine learning does not always provide significant improvements over classical statistical methods in clinical prediction tasks [6]. Nevertheless, such studies have not considered the impact of feature homogeneity. Feature homogeneity may reduce class separability and the discriminative capability of models. Furthermore, investigations into the effectiveness of gradient boosting methods in clinical prediction have primarily focused on performance optimization within a single dataset. They have not examined generalization ability across multiple datasets [7]. A similar limitation is observed in ensemble learning approaches. These approaches have been shown to improve prediction stability but are still evaluated within a single dataset context [8]. On the other hand, research on feature selection for Naïve Bayes [9] demonstrates potential performance improvements. However, it does not assess the impact under conditions of feature dependency or within multi-dataset scenarios. Overall, these studies remain focused on single-dataset benchmarking and local performance optimization. Thus, they do not provide sufficient insight into model robustness against distributional changes (dataset shift) within the same clinical domain.

Despite numerous studies, significant gaps in literature remain. Most research assumes datasets have enough feature variability. As a result, the impact of feature homogeneity on model performance is largely unexamined. Robustness is usually evaluated on a single dataset, without considering distributional changes across datasets in the same clinical domain (dataset shift). From a dataset shift and distributional robustness perspective, most prior studies assume training and testing data are drawn from identical or very similar distributions. However, in real-world clinical practice, this rarely happens. Data collection protocols, patient characteristics, and feature aggregation levels differ across healthcare systems [10], [11]. This leads to a significant research gap. There is no systematic evaluation of how machine learning models behave when applied to other clinical datasets with different feature structures in the same domain. Also, the influence of feature homogeneity on model robustness is rarely explored in cross-dataset settings. Studies show that dataset shift can degrade model performance in clinical applications [12]. This highlights the need to evaluate robustness under changing data distributions. Formally, dataset shift happens when training and testing data distributions aren't identical. This is a major challenge in developing reliable, generalizable models [13].Unlike previous studies that mainly focus on single-dataset benchmarking and local performance optimization, this study explicitly evaluates cross-dataset robustness across clinical datasets and investigates the influence of feature homogeneity as a key factor affecting model stability under distributional shift.

The objective of this study is to evaluate the robustness of machine learning and deep learning models, including ensemble methods, in patient adherence classification. It also aims to compare model performance across datasets with different structural characteristics. Additionally, the study analyzes the impact of feature 99iiiii0069engineering, hyperparameter tuning, and feature extraction on model stability. Overall, this study demonstrates that feature homogeneity significantly affects a model's ability to distinguish between classes and emphasizes the importance of selecting models that align with dataset characteristics to achieve stable and reliable predictions in clinical applications. Furthermore, this research highlights that the robustness of models on tabular clinical data is not solely determined by algorithmic complexity, but is also strongly influenced by dataset structure, feature homogeneity, and distributional shifts across datasets.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 79-92

## 2. Material and Methods

This study proposes an AutoML-based classification framework to evaluate model robustness across clinical datasets with different structural characteristics. Robustness is the model's ability to maintain consistent performance despite differences in class distributions, feature heterogeneity, and population origins. The pipeline integrates preprocessing, feature engineering, feature selection, dimensionality reduction, data balancing, and model optimization in a unified workflow. Model performance is evaluated using accuracy, precision, recall, F1-score, and consistency through repeated stratified cross-validation and cross-dataset testing. This design supports reproducibility and enables systematic analysis of the impact of dataset structure and preprocessing strategies on model performance.

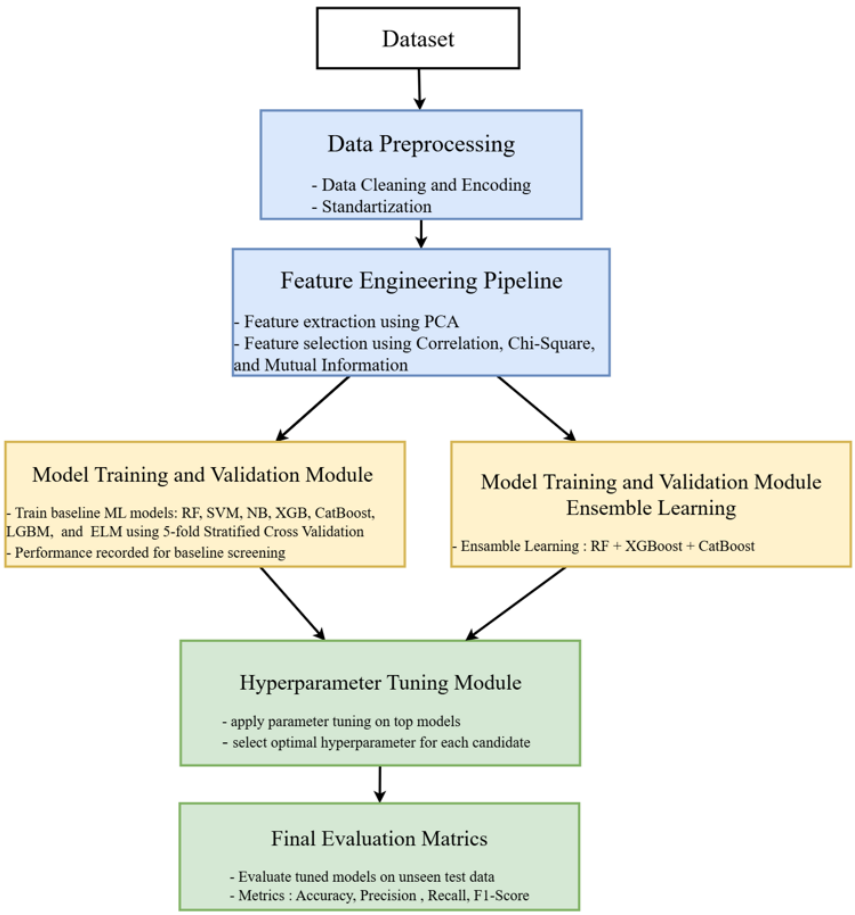


**Fig. 1** *Flow diagram*

**Algorithm 1:** *Automated Machine Learning Pipeline for Classification with Feature Engineering and Hyperparameter Optimization*

Begin

1. Load dataset D and define target y
2. Split D into training (D_train) and testing (D_test)
3. Preprocess D_train:
   - Handle missing values
   - Encode categorical features
   - Normalize numerical features
4. Perform feature engineering:
   - Generate interaction, clustering, and time-based features
5. Apply feature selection:
   - Remove low-variance and highly correlated features
   - Select important features (Boruta / RFE)

6. Reduce dimensionality using PCA (95% variance)
7. If data is imbalanced, apply SMOTE
8. For each candidate model Mi:
   - Perform hyperparameter tuning
   - Train and evaluate using cross-validation
   - Store performance metrics
9. Select best model X based on evaluation metrics
10. Evaluate X on D_test
11. Return X and performance results

End

## 2.1 Dataset

This study employs two clinical datasets to evaluate model robustness under varying data characteristics: a patient adherence dataset and a health risk claim dataset, both sourced from Kaggle, as summarized in Table 1. The patient adherence dataset is publicly available on Kaggle http://kaggle.com/datasets/vipulshahi/patient-adherence-dataset, ensuring accessibility and reproducibility. It contains 5,004 samples and 18 features, including demographic, clinical, socio-economic, and behavioral attributes, with a binary target (adherent vs. non-adherent). This dataset exhibits a relatively high proportion of missing values 22.21% and a near-balanced class distribution imbalance ratio: 1.19. The second dataset, referred to as the health risk claims dataset, is also obtained from Kaggle and available at https://www.kaggle.com/datasets/dd353efc2f392e03a781933a42c0e326ceda8639cfa21874990e16b7aa6f65d1?resource=download. It consists of 8,141 samples and 17 features, with similar attribute categories but derived from a different population. In contrast to the first dataset, it contains no missing values 0.00% and is fully balanced imbalance ratio: 1.00, providing a more stable setting for evaluating model generalization. The descriptive statistics in Table 1 include the number of samples, feature count, proportion of missing values, imbalance ratio, and average inter-feature correlation. The patient adherence dataset shows a low correlation 0.027, indicating limited linear redundancy, whereas the health claims dataset exhibits a higher correlation 0.174, suggesting stronger feature relationships. PCA-based visualization further reveals greater class overlap in the adherence dataset and clearer separation in the claim's dataset. To ensure consistency and full reproducibility of the experiments, both datasets undergo standardized preprocessing steps, including missing value imputation, categorical encoding, and feature scaling (normalization and standardization), prior to model training and evaluation. All datasets are publicly accessible and can be directly downloaded using the provided links, enabling full replication of the study.

**Table 1** *Dataset configuration*

| Dataset | Samples | Feature | Missing | Imbalance Ratio | Correlation |
|---|---|---|---|---|---|
| Patient Adherence Dataset | 5.004 | 18 | 22.21% | 1.19 | 0.027 |
| Health Claims Dataset | 8.141 | 17 | 0.00% | 1.00 | 0.174 |

## 2.2 Feature Engineering Pipeline

An integrated feature engineering pipeline was applied to enhance the predictive performance of machine learning and deep learning models [14] .The pipeline consists of two main stages: feature extraction and feature selection. In the feature extraction stage, numerical features are transformed using natural logarithms and square functions. This approach captures potential non-linear relationships, as expressed in Equation (1).

$$x_{log} = log\ (1 + x), x_{square} = x^2 \tag{1}$$

Where $x_{log}$ dampens large values, $x_{square}$ emphasizes large values, captures quadratic/non-linear growth. Meanwhile, categorical features were transformed using category occurrence frequency, as expressed in Equation (2).

$$x_{freq} = count(x_i) \tag{2}$$

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 79-92

After the initial transformations, Principal Component Analysis (PCA) was applied to reduce dimensionality and extract the principal components that retain the largest variance in the dataset. PCA projects the original data $X \in \mathbb{R}^{n\times p}$into a new feature space $Z \in \mathbb{R}^{n\times k}$ through a linear transformation, as shown in Equation (3).

$$Z = XW \tag{3}$$

where $W \in \mathbb{R}^{p\times k}$ is the matrix of eigenvectors obtained from the eigen decomposition of the covariance matrix. $\Sigma = \frac{1}{n-1}X^T X$. Although the patient adherence dataset used in this study exhibits a high degree of feature homogeneity, PCA was still applied for exploratory and preventive purposes. Specifically, the application of PCA aims to:

- Reduce redundancy among statistically highly correlated features,
- Improve the numerical stability of models operating in high-dimensional feature spaces, and
- Evaluate the impact of dimensionality reduction on model performance in homogeneous datasets compared to more heterogeneous datasets [15].

Principal Component Analysis (PCA) is applied using a variance-based approach, retaining components that explain 95% of the cumulative explained variance, resulting in an adaptive number of components that depend on the data characteristics and are reported in the experimental results. The feature selection process is conducted through a multi-stage pipeline. First, features with near-zero variance are removed using a Variance Threshold of $1\times10^{-6}$. Second, feature redundancy is reduced using a Pearson correlation filter with a threshold of 0.98. Subsequently, model-based feature selection is performed using Boruta (Random Forest, maximum 50 iterations), with a fallback to Select from Model if necessary, and finalized using Recursive Feature Elimination (RFE) to limit the final feature set to 30 features. For categorical features, high-cardinality variables undergo target encoding (smoothing = 10.0), while the remaining features use One-Hot Encoding. Feature engineering also includes interaction-only Polynomial Features and clustering-based features from K-Means with an adaptive number of clusters. Chi-Square and Mutual Information are not used, as model-based feature selection is preferred for capturing non-linear feature relationships. The impact of PCA on model performance is quantitatively evaluated by comparing results before and after transformation using metrics such as accuracy, precision, recall, F1-score, and AUC. These results are presented in the experimental section to demonstrate PCA's contribution to model stability and generalization performance. The methods applied include:

a. Pearson correlation [16], with the Equation (4)

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \tag{4}$$

Where $r_{xy}$ pearson correlation coefficient between variables X and Y, $x_i$ is i-th value of variable X, $y_i$ is i-th value of variable Y, $\bar{x}$ is mean value of variable X, $\bar{y}$ is mean value of variable Y.

b. Chi-Square ($\chi^2$) [15] test for discrete features, with the Equation (5).

$$\chi^2 = \sum \frac{(O_i-E_i)^2}{E_i} \tag{5}$$

Where $\chi^2$ is Chi-Square test statistic, $O_i$ is Observed frequency for category i, $E_i$ is Expected frequency for categories i.

c. Mutual Information (MI) [15] to measure the dependency between features and the target, with the equation (6)

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p\,(x,y) log\ \frac{p(x,y)}{p(x)p(y)} \tag{6}$$

Where $I(X;Y)$ is Mutual information between variables X and Y, X, Y is Random variables X and Y, $x \in X$ is possible values of variable X, $y \in Y$ is possible values of variable Y, $p(x)$ is marginal probability of x, $p(y)$ is marginal probability of y, $p(x,y)$ is probability of x and y. The results of this multi-criteria approach enable the selection of the most relevant features. This process reduces redundancy and less informative features. As a result, model stability and discriminative ability are improved, especially on limited clinical datasets with low separability.

## 2.3 Machine Learning and Deep Learning Models

In this study, eight classification approaches were tested: Random Forest, SVM, Naive Bayes, XGBoost, CatBoost, LightGBM, ensemble models, and ELM. The selection of these algorithms was based on the characteristics of tabular clinical datasets, including mixed numerical and categorical features, potential nonlinearity, and the need for a balance of reliability, interpretability, and flexibility.

a. **Random Forest**
   This method is referred to as a bagging-based ensemble, as it is known for being stable against noise and overfitting [17], capable of handling both numerical and categorical data, and for providing feature importance estimates, thereby aiding interpretability [18].
b. **Support Vector Machine**
   This method is suitable for high-dimensional data and can find an optimal hyperplane for class separation [19]. It is especially effective when the margin between classes is clear, and the dataset is not too large [20]. The combination of kernels enables it to handle non-linearity.
c. **Naïve Bayes**
   A fast, robust probabilistic method, effective with high-dimensional data and when the feature-independence assumption holds or with simple data structures [21] [22].
d. **XGBoost**
   A popular gradient boosting method that captures complex relationships, reduces overfitting via regularization, delivers computational efficiency, and typically performs well in medical and risk classification tasks [23].
e. **CatBoost**
   This gradient boosting variant handles categorical features directly, leverages ordered boosting and effective regularization and is well-suited to datasets with many categorical features or high heterogeneity [24] [25].
f. **LightGBM**
   A boosting alternative optimized for speed and memory efficiency, handling large-scale tabular data with a strong balance of training speed and accuracy [26] [27].
g. **Ensemble Learning**
   In this study, the ensemble learning approach combined predictions from Random Forest, XGBoost, and CatBoost models. This method was employed to improve overall predictive accuracy and robustness. By aggregating the strengths of each base model, the ensemble aims to reduce individual model biases and better generalize to diverse clinical data.
h. **Extreme Learning Machine**
   Ted A lightweight neural network model with a fast-training procedure (requiring only a single pass without backpExtreme Learning Machine is a lightweight neural network model featuring a fast-training procedure requiring only a single pass without backpropagation iterations. It is suitable for small- to medium-sized datasets and environments with limited computational resources [28].

## 2.4 Experiment and Evaluation

Stratified 5-fold cross-validation with 10 repetitions is used to obtain robust, reliable estimates of model performance under real-world conditions. Stratification preserves the original class distribution in each fold. This preservation is essential for handling class imbalance in clinical datasets. In each iteration, one-fold is used as the test set, and the remaining folds are used for training. This procedure yields a more stable and less biased evaluation of model performance. A fixed random seed of 42 is applied across all experiments to ensure reproducibility. The seed is consistently used for data splitting, model initialization, and evaluation. The experiments use Python 3.14.0 and the following key libraries: NumPy 1.26.4, Pandas 2.2.2, Scikit-learn 1.4.2, XGBoost 3.1.2, CatBoost 1.2.8, LightGBM 4.6.0, and HPELM 1.0.10. These dependencies ensure a consistent and reproducible computational environment. Model performance is influenced by both learned parameters and hyperparameters such as tree depth, number of estimators, learning rate, regularization strength, and kernel configuration. The hyperparameter search space is large and complex, so a systematic Hyperparameter Optimization (HPO) approach is adopted [29]. Randomized Search Cross-Validation (RandomizedSearchCV) is chosen for its efficiency in exploring large hyperparameter spaces. The defined hyperparameter search space is as follows:

- Random Forest: number of estimators ∈ {100, 200, 400}, max depth ∈ {None, 5, 10}
- XGBoost: learning rate ∈ [0.01, 0.1], max depth ∈ {3, 5}, number of estimators ∈ {50, 100, 200}
- LightGBM: number of leaves ∈ {15, 31}, learning rate ∈ [0.01, 0.1]
- SVM: regularization parameter C ∈ {0.1, 1, 10}, kernel = RBF
- Logistic Regression: C ∈ {0.01, 0.1, 1, 10}.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 79-92

The optimization process is integrated with cross-validation, which involves repeatedly splitting the data and training the model on one part while validating it on another. This reduces the risk of overfitting, where a model performs well on training data but poorly on new data. All models are evaluated using various performance metrics: accuracy (the overall correctness), precision (correct positive predictions out of all positive predictions), recall (correct positive predictions out of all actual positives), F1-score (the harmonic mean of precision and recall), and AUC (the area under the Receiver Operating Characteristic curve, measuring the trade-off between true positive and false positive rates). Together, these provide a comprehensive assessment of model performance.

## 2.5 Performance Evaluation Matrices

In this study, five performance evaluation metrics are used to assess the effectiveness of the model, namely:

**a. Accuracy**

Accuracy measures the proportion of correctly predicted instances relative to all predictions made by the model [15], as expressed in Equation (7)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Where TP refers to True Positive, TN to True Negative, FP to False Positive, and FN to False Negative. Accuracy is appropriate when the dataset is relatively balanced; however, it can be misleading in imbalanced datasets because majority-class predictions may dominate the accuracy score.

**b. Precision**

Precision measures how accurately the model predicts the positive class. Precision indicates the proportion of positive predictions that are truly positive [15], as shown in Equation (8).

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

Where TP is True Positive, FP is False Positive, and FN is False Negative.

**c. Recall**

Recall evaluates the model's ability to detect all actual positive cases [15]. The higher the recall, the fewer positive cases are missed by the model, as shown in Equation (9).

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

Where TP is True Positive, FP is False Positive, and FN is False Negative.

**d. F1-Score**

The F1-Score is the harmonic means of precision and recall, providing a balanced measure between the two metrics. The F1-Score is particularly useful when the dataset is imbalanced, as it considers both precision and recall simultaneously [15], as shown in Equation (10).

$$F1 - Score = 2\,x\,\frac{Precision+Recall}{Precision+Recall} \tag{10}$$

A high F1-Score indicates that the model is not only accurate in predicting positive cases (high precision), but also successfully identifies most of the actual positive cases (high recall).

## 3. Result and Discussion

The model performance evaluation was conducted to classify patient adherence levels using the Kaggle patient adherence dataset and the health risk claims dataset. The review was carried out in three stages: baseline performance evaluation without hyperparameter tuning, advanced performance evaluation after hyperparameter tuning, and performance evaluation using a feature engineering pipeline (feature extraction and feature selection). This three-stage approach aims to determine how well each model's initial configuration learns from the data and how parameter optimization can improve classification performance across different robustness conditions. All models were tested using a 5-fold cross-validation with 10 repetitions to ensure stable, unbiased results.

## 3.1 Results of Model Evaluation Without Hyperparameter Tuning

The baseline results in Table 2 show that performance differences among models are relatively narrow. Accuracy ranges from 0.5816 to 0.6232, and variability is low (0.0036–0.0077). The small standard deviations reflect this limited variability. The Wilcoxon signed-rank test also indicates that the differences between models are not statistically significant. This suggests that the data set is stable with limited discriminative power. As a result, performance improvements are mostly constrained by the underlying data structure. CatBoost (0.6232 ± 0.0036) and SVM (0.6227 ± 0.0043) are the top-performing models. These models only slightly outperform Random Forest and Extreme Learning Machine (ELM). Ensemble and boosting methods do not provide meaningful advantages. In contrast, Naive Bayes achieves the lowest performance (0.5816 ± 0.0319) and has higher variance. This result suggests that the independence assumption is poorly satisfied and that more flexible decision boundaries are needed. Overall, no single model shows a dominant advantage. The dataset likely has limited nonlinear complexity. As a result, ensemble-based approaches offer little benefit. Combining models does not significantly improve performance over individual learners. Instead, model behavior is mainly influenced by data structure, not by algorithmic complexity. The observed convergence in performance is also consistent with prior literature. This literature emphasizes that machine learning performance depends largely on data representation, preprocessing quality, and feature engineering. The inherent superiority of any algorithm is less important. Therefore, improving data quality and feature design will likely have a greater impact on performance than choosing more complex models [30].

**Table 2** *Average model performance before hyperparameter tuning*

| Model | Average accuracy | Standard deviation | Precision | Standard deviation | Recall | Standard deviation | F1-score | Standard deviation |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 0.6195 | ±0.0065 | 0.6181 | ±0.0065 | 0.6195 | ±0.0065 | 0.6173 | ±0.0064 |
| Support Vector Machine | 0.6227 | ± 0.0043 | 0.6231 | ±0.0048 | 0.6227 | ±0.0043 | 0.6222 | ±0.0045 |
| Naive Bayes | 0.5816 | ± 0.0319 | 0.5605 | ±0.0709 | 0.5816 | ±0.0319 | 0.5069 | ±0.0841 |
| XGBoost | 0.5953 | ± 0.0046 | 0.5941 | ±0.0047 | 0.5953 | ±0.0046 | 0.5940 | ±0.0045 |
| CatBoost | 0.6232 | ± 0.0036 | 0.6222 | ±0.0037 | 0.6232 | ±0.0036 | 0.6219 | ±0.0036 |
| LightGBM | 0.6134 | ± 0.0062 | 0.6126 | ±0.0065 | 0.6134 | ±0.0062 | 0.6123 | ±0.0065 |
| Extreme Learning Machine | 0.6187 | ± 0.0058 | 0.6179 | ±0.0058 | 0.6187 | ±0.0058 | 0.6172 | ±0.0058 |
| Ensemble Learning (Random Forest, XG Boost dan Cat Boost) | 0.6147 | ± 0.0077 | 0.6135 | ±0.0079 | 0.6147 | ±0.0077 | 0.6131 | ±0.0078 |

## 3.2 Results of Model Evaluation Using Hyperparameter Tuning Without the Feature Engineering Pipeline

The results in Table 3 show that after hyperparameter tuning, improvements across models are within a narrow range with small absolute differences. Random Forest achieved the highest accuracy (0.6326 ± 0.0047). XGBoost (0.6296 ± 0.0036) and the ensemble model (0.6292 ± 0.0132) followed closely. Compared to the previous baseline range of 0.58 to 0.62, tuning gives marginal yet consistent improvements, especially for tree-based ensemble models. This pattern suggests that non-linear structures and complex feature interactions are key in this dataset. However, the combined ensemble of Random Forest, XGBoost, and CatBoost does not significantly beat the best single model. This indicates high correlation among its parts, which limits the benefit of diversity. In contrast, Naive Bayes performs the worst and has high variability (standard deviation up to 0.0841 in F1-score). This confirms that the independence assumption does not fit the data. From an analytical perspective, XGBoost and Random Forest are more stable and exhibit more consistent generalization than methods such as Naive Bayes or Extreme Learning Machine. Statistical tests using the Wilcoxon signed-rank test show that performance differences among models are not statistically significant ($p > 0.05$). Thus, observed numerical advantages are insufficient to claim any model's superiority. This supports that all models perform similarly on this dataset. Potential biases also need attention, such as non-representative data or class imbalance. These can skew metrics like accuracy and recall. The larger standard deviation in the ensemble model suggests it is sensitive to data splitting, potentially affecting reproducibility.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 79-92

### 3.3 Evaluation Model Using Feature Engineering Pipeline and Hyperparameter Tuning

The integration of a feature engineering pipeline, Principal Component Analysis (PCA), and hyperparameter tuning yields different performance patterns across the two datasets used: the Kaggle patient adherence dataset and a health-risk claim dataset from previous research. For a comprehensive analysis, results are compared with the baseline in Table 1 and with tuning-only (without feature engineering) in Table 3, as summarized in Tables 5 and 6. In this study, PCA is applied for dimensionality reduction, retaining 95% of the explained variance. This reduces feature redundancy, addresses potential multicollinearity, and preserves key information in principal components. In Table 4, for the patient adherence dataset, feature engineering combined with PCA yields only limited improvement. Model accuracy remains within a narrow range (0.5614–0.5703), showing that dimensionality reduction does not greatly enhance class separability. This suggests the dataset is relatively simple, with the main information already well represented by original features and lacking strong nonlinear relationships. Thus, Naive Bayes and XGBoost still achieve competitive results, indicating PCA's contribution is minimal. In contrast, Table 5, using the health-risk claim dataset, shows feature engineering and PCA have a much greater impact. Boosting models such as CatBoost (0.8333), LightGBM (0.8305), and XGBoost (0.8253) improved substantially over baseline results in Table 2, which were around 0.62. This indicates PCA successfully reduces noise and improves separability, allowing models to better capture the data's complex patterns. Compared with prior results in Table 3, the improvement is also clear. Random Forest rises from 0.6169 to 0.8292, and SVM jumps from 0.5433 to 0.8137. These results show that performance gains are driven mainly by improved data representation from feature engineering and PCA, rather than by algorithm choice alone. To confirm these differences, a Wilcoxon signed-rank test was conducted. For the health-risk claim dataset, performance differences among baseline models, tuning-only models, and PCA-based feature engineering are statistically significant ($p < 0.05$). However, for the patient adherence dataset, the differences are not statistically significant ($p > 0.05$). This confirms that PCA's effectiveness depends on the dataset structure complexity. Boosting-based algorithms such as CatBoost, LightGBM, and XGBoost perform best on complex data by effectively leveraging PCA-transformed features. By contrast, Naive Bayes is better for simpler datasets where features are more independent.

**Table 3** *Average model performance after applying hyperparameter tuning*

| Model | Average accuracy | Standard deviation | Precision | Standard deviation | Recall | Standard deviation | F1-score | Standard deviation |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 0.6326 | ± 0.0047 | 0.6319 | ± 0.0049 | 0.6326 | ± 0.0047 | 0.6299 | ± 0.0048 |
| Support Vector Machine | 0.6247 | ± 0.0048 | 0.6275 | ± 0.0057 | 0.6247 | ± 0.0048 | 0.6239 | ± 0.0056 |
| Naive Bayes | 0.5816 | ± 0.0319 | 0.5605 | ± 0.0709 | 0.5816 | ± 0.0319 | 0.5069 | ± 0.0841 |
| XGBoost | 0.6296 | ± 0.0036 | 0.6290 | ± 0.0038 | 0.6296 | ± 0.0036 | 0.6283 | ± 0.0038 |
| CatBoost | 0.6241 | ± 0.0044 | 0.6233 | ± 0.0044 | 0.6241 | ± 0.0044 | 0.6229 | ± 0.0043 |
| LightGBM | 0.6228 | ± 0.0077 | 0.6221 | ± 0.0080 | 0.6228 | ± 0.0077 | 0.6199 | ± 0.0078 |
| Extreme Learning Machine | 0.6191 | ± 0.0065 | 0.6183 | ± 0.0066 | 0.6191 | ± 0.0065 | 0.6177 | ± 0.0066 |
| Ensemble Learning (Random Forest, XG Boost dan Cat Boost) | 0.6292 | ± 0.0132 | 0.6279 | ± 0.0133 | 0.6292 | ± 0.0132 | 0.6281 | ± 0.0133 |

**Table 4** *Average model accuracy using feature extraction on the patient adherence dataset*

| Model | Average accuracy | Standard deviation | Precision | Standard deviation | Recall | Standard deviation | F1-score | Standard deviation |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 0.5684 | ± 0.0061 | 0.5648 | ± 0.0067 | 0.5684 | ± 0.0061 | 0.5530 | ± 0.0081 |
| Support Vector Machine | 0.5673 | ± 0.0061 | 0.5642 | ± 0.0067 | 0.5673 | ± 0.0061 | 0.5516 | ± 0.0140 |
| Naive Bayes | 0.5614 | ± 0.0057 | 0.5598 | ± 0.0051 | 0.5614 | ± 0.0057 | 0.5545 | ± 0.0035 |
| XGBoost | 0.5703 | ± 0.0070 | 0.5669 | ± 0.0073 | 0.5703 | ± 0.0070 | 0.5596 | ± 0.0102 |
| CatBoost | 0.5668 | ± 0.0055 | 0.5629 | ± 0.0060 | 0.5668 | ± 0.0055 | 0.5542 | ± 0.0064 |
| LightGBM | 0.5677 | ± 0.0059 | 0.5643 | ± 0.0064 | 0.5677 | ± 0.0059 | 0.5425 | ± 0.0110 |
| Extreme Learning Machine | 0.5635 | ± 0.0030 | 0.5584 | ± 0.0038 | 0.5635 | ± 0.0030 | 0.5409 | ± 0.0103 |
| Ensemble Learning (Random Forest, XG Boost dan Cat Boost) | 0.5669 | ± 0.0061 | 0.5632 | ± 0.0062 | 0.5669 | ± 0.0061 | 0.5546 | ± 0.0064 |

## 3.4 Analysis of Model Robustness

To evaluate the generalization ability of the models under distributional shift, robustness is defined in this study as the relative performance degradation between two datasets with different complexity levels, namely the patient adherence dataset and the health-risk claim dataset. Unlike a simple accuracy difference, robustness is formally defined as a normalized measure as follows Equation (11)

$$\text{Robustness} = \left(\frac{A_{risk} - A_{patient}}{A_{risk}}\right) \times 100\% \tag{11}$$

where $A_{risk}$ represents the model accuracy on the health-risk claim dataset and $A_{patient}$ represents the model accuracy on the patient adherence dataset. This formulation reflects the proportion of performance change relative to the more complex dataset, providing a more standardized interpretation of model sensitivity to distributional shifts. Based on the robustness results in Table 6, indicate that ensemble and boosting-based models such as Random Forest, XGBoost, CatBoost, and LightGBM demonstrate consistently strong generalization ability under dataset shift conditions. Their robustness values, which range from approximately 30% to 32%, indicate relatively stable performance when transitioning from the simpler patient adherence dataset to the more complex health-risk claim dataset. XGBoost and CatBoost achieve the highest robustness values at 31.98%, closely followed by LightGBM 31.64% and Random Forest 31.46%, confirming their effectiveness in maintaining predictive stability across different data distributions. In contrast, the Naive Bayes model exhibits a substantial negative robustness value -67.33%, indicating a severe degradation in performance when applied to the more complex dataset. This result highlights the limitation of the conditional independence assumption in Naive Bayes, especially in scenarios involving strong feature dependencies and nonlinear relationships. Similarly, although Support Vector Machine and Extreme Learning Machine maintain positive robustness values (around 30%), their performance remains slightly lower compared to boosting-based approaches, suggesting moderate sensitivity to dataset complexity. Interestingly, the Ensemble Learning model (combining Random Forest, XGBoost, and CatBoost) achieves a lower robustness value 24.73% compared to individual boosting models. This suggests that the ensemble strategy used may not fully exploit the complementary strengths of its base learners, potentially due to suboptimal weighting or integration mechanisms. Overall, the robustness evaluation demonstrates that model stability is strongly influenced by the ability to capture complex and nonlinear feature interactions. Boosting-based models exhibit superior robustness due to their flexibility in handling heterogeneous feature spaces, while simpler probabilistic models such as Naive Bayes are highly sensitive to changes in feature dependency structures.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 79-92

**Table 5** *Average model accuracy using feature extraction on the health risk claim dataset*

| Model | Average accuracy | Standard deviation | Precision | Standard deviation | Recall | Standard deviation | F1-Score | Standard deviation |
|---|---|---|---|---|---|---|---|---|
| Random Forest | 0.8292 | ± 0.0030 | 0.8053 | ± 0.0045 | 0.8292 | ± 0.0030 | 0.8060 | ± 0.0040 |
| Support Vector Machine | 0.8137 | ± 0.0011 | 0.7783 | ± 0.0058 | 0.8137 | ± 0.0011 | 0.7437 | ± 0.0031 |
| Naive Bayes | 0.3356 | ± 01165 | 0.8072 | ± 0.0118 | 0.3356 | ± 0.1165 | 0.2603 | ± 0.1519 |
| XGBoost | 0.8253 | ± 0.0028 | 0.8069 | ± 0.0036 | 0.8253 | ± 0.0028 | 0.8119 | ± 0.0032 |
| CatBoost | 0.8333 | ± 0.0030 | 0.8120 | ± 0.0042 | 0.8333 | ± 0.0030 | 0.8127 | ± 0.0038 |
| LightGBM | 0.8305 | ± 0.0037 | 0.8114 | ± 0.0046 | 0.8305 | ± 0.0037 | 0.8152 | ± 0.0042 |
| Extreme Learning Machine | 0.8109 | ± 0.0005 | 0.7502 | ± 0.0412 | 0.8109 | ± 0.0005 | 0.7277 | ± 0.0010 |
| Ensemble Learning (Random Forest, XG Boost dan Cat Boost) | 0.8160 | ± 0.0053 | 0.7993 | ± 0.0058 | 0.8160 | ± 0.0053 | 0.8051 | ± 0.0054 |

**Table 6** *Model robustness*

| Model | Patient adherence accuracy | Risk claims accuracy | Robustness (%) |
|---|---|---|---|
| Random Forest | 0.5684 | 0.8292 | 31.46% |
| Support Vector Machine | 0.5673 | 0.8137 | 30.28% |
| Naive Bayes | 0.5614 | 0.3356 | -67.33% |
| XGBoost | 0.5703 | 0.8253 | 31.98% |
| CatBoost | 0.5668 | 0.8333 | 31.98% |
| LightGBM | 0.5677 | 0.8305 | 31.64% |
| Extreme Learning Machine | 0.5635 | 0.8109 | 30.51% |
| Ensemble Learning (Random Forest, XG Boost dan Cat Boost) | 0.5669 | 0.8160 | 24.73% |

## 3.5 Discussion

The comprehensive evaluation across baseline, hyperparameter tuning, feature engineering, and robustness analysis demonstrates that model performance is primarily determined by dataset complexity and feature representation, rather than by algorithmic sophistication. On the patient adherence dataset, all models exhibit narrow, statistically non-significant performance differences, indicating limited intrinsic separability and a weak nonlinear structure. Hyperparameter tuning yields only marginal improvements, confirming that parameter optimization alone is insufficient without enhancing feature representation [31]. This convergence further suggests that most models operate within similar decision boundaries, reflecting their low feature complexity. The effect of PCA-based feature engineering further highlights the role of data structure. On the patient adherence dataset, PCA provides negligible improvement. This shows that the original feature space already captures most discriminative information with minimal redundancy. In contrast, on the more complex health-risk claim dataset, PCA significantly improves performance. This effect is particularly strong for boosting-based models (such as CatBoost, XGBoost, and LightGBM). High-complexity, interaction-rich datasets benefit more from dimensionality reduction, which enhances the signal-to-noise ratio. This approach also facilitates nonlinear pattern learning by simplifying the geometry of the feature space [32]. The sharp performance degradation of Naive Bayes on the health-risk claims dataset accuracy 0.3356 provides strong theoretical evidence of violations of the conditional independence assumptions. In complex feature spaces, strong inter-feature dependencies invalidate the factorization of joint probability distributions, leading to biased posterior estimation and unstable classification behavior [33]. This structural limitation explains its severe performance collapse under high-feature-interaction conditions [34]. In contrast, boosting-based models consistently outperform because they can iteratively model higher-order feature interactions via hierarchical decision trees, enabling explicit learning of nonlinear dependencies without relying on independence assumptions [35]. Robustness analysis further reinforces these findings under distributional shift conditions. Boosting-based models maintain stable robustness values (30–32%), reflecting strong generalization across heterogeneous datasets. In contrast, Naive Bayes exhibits a highly

negative robustness value (-67.33%), indicating extreme sensitivity to changes in feature distribution and dependency structure. Notably, the ensemble model (Random Forest, XGBoost, CatBoost) shows lower robustness (24.73%) compared to individual boosting models, suggesting that high inter-model correlation limits diversity and reduces stability gains. Overall, these results confirm that model effectiveness is fundamentally determined by the alignment between algorithmic assumptions and data characteristics: probabilistic models are suitable for low-interaction settings, whereas boosting-based ensembles are superior for complex, nonlinear, and dependency-rich data.

## 4. Conclusions

This study shows that classification model performance and robustness depend more on dataset characteristics than algorithmic complexity. On simple, homogeneous datasets, all models perform similarly, so added model complexity provides little benefit. On complex, heterogeneous datasets, boosting-based models excel at capturing nonlinear patterns and feature interactions. These findings highlight that a mismatch between model assumptions and data structure, particularly in models such as Naive Bayes, can lead to substantial performance degradation under strong feature interaction conditions. Furthermore, cross-dataset evaluation indicates that model stability largely depends on its ability to adapt to distributional changes and boosting-based models consistently demonstrate greater robustness than simpler probabilistic approaches. The main contribution of this study is to demonstrate that feature homogeneity and data complexity are key factors influencing both performance and robustness, surpassing the impact of algorithm selection alone. Therefore, model selection in clinical applications should explicitly consider data characteristics rather than relying solely on model complexity. This study is limited to two datasets from the same domain. It does not use domain adaptation techniques. Future research should explore cross-domain datasets and develop more adaptive methods to handle variations in data distribution.

## Acknowledgement

The authors express their gratitude to the Artificial Intelligence Innovation Centre at the University of Brawijaya for providing laboratory analysis and invaluable support.

## Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

## Author Contribution

*The authors confirm contribution to the paper as follows:* ***study conception and design:*** *Wayan Firdaus Mahmudy, Silfiana Nur Hamida;* ***data collection:*** *Silfiana Nur Hamida, Wayan Firdaus Mahmudy;* ***analysis and interpretation of results:*** *Silfiana Nur Hamida, Wayan Firdaus Mahmudy;* ***draft manuscript preparation:*** *Silfiana Nur Hamida, Wayan Firdaus Mahmudy. All authors reviewed the results and approved the final version of the manuscript.*

## References

[1] U. Religioni, R. B. Rodríguez, . P. Requena, M. Borowska and . J. Ostrowski, "Enhancing Therapy Adherence: Impact on Clinical Outcomes, Healthcare Costs, and Patient Quality of Life," *Medicina,* vol. 61, no. 1, p. 153, 2025, doi : https://doi.org/10.3390/medicina61010153.

[2] J. Pasaribu, N. Yudistira and W. F. Mahmudy, "Tabular Data Classification and Regression: XGBoost or Deep Learning With Retrieval-Augmented Generation," *IEEEAccess,* vol. 12, pp. 191719-191732, 2024, doi : https://doi.org/10.1109/access.2024.3518205.

[3] C. D. Marineci , A. Valeanu, C. Chirita, S. Nigres, C. Stoicescu and V. Chioncel, "Development and Validation of Predictive Models for Non-Adherence to Antihypertensive Medication," *Medicine,* vol. 61, no. 7, p. 1313, 2025, doi: https://doi.org/10.3390/medicina61071313.

[4] F. Orhan and M. N. Kurutkan, "Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors," *BMC Health Services Research,* vol. 25, p. 366, 2025, doi : https://doi.org/10.1186/s12913-025-12502-5.

[5] S. A. Fayaz, M. Zaman, S. Kaul and M. A. Butt, "Is Deep Learning on Tabular Data Enough? An Assessment," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 13, no. 4, pp. 466 - 473, 2022, doi: https://dx.doi.org/10.14569/IJACSA.2022.0130454.

[6] A. Cerasa, G. Tartarisco, R. Bruschetta, I. Ciancarelli, G. Morone, R. S. Calabrò, G. Pioggia, P. Tonin and M. Iosa, "Predicting Outcome in Patients with Brain Injury: Differences between Machine Learning versus Conventional Statistics," *biomedicines,* vol. 10, no. 2267, pp. 1-15, 2022, doi: https://doi.org/10.3390/biomedicines10092267.

[7] S. M. Ganie, P. K. D. Pramanik, M. B. Malik, . S. Mallik and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Frontiers in Genetics,* vol. 14, 2023, doi: https://doi.org/10.3389/fgene.2023.1252159.

[8] A. Alotaibi, "Ensemble Deep Learning Approaches in Health Care: A Review," *Computers, Materials and Continua,* vol. 82, no. 3, pp. 3741-3771, 2025, doi; https://doi.org/10.32604/cmc.2025.061998.

[9] E. O. Omuya, G. O. Okeyo and M. W. Kimwele, "Feature Selection for Classification using Principal Component Analysis and Information Gain," *Expert Systems with Applications,* vol. 174, p. 114765, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114765.

[10] G. F. d. S. Silva, F. N. B. Filho, R. M. Wichmann, F. C. d. S. Junior and A. D. P. C. Filho, "Strategies for detecting and mitigating dataset shift in machine learning for health predictions: A systematic review," *Journal of Biomedical Informatics,* vol. 170, p. 104902, 2025, doi: https://doi.org/10.1016/j.jbi.2025.104902.

[11] L. Han, "Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings," *JAMA Network Open,* vol. 8, no. 6, p. 2513705, 2025, doi: https://doi.org/10.1001/jamanetworkopen.2025.13705.

[12] S. Lee, C. Yin and P. Zhang, "Stable clinical risk prediction against distribution shift in electronic health records," *Patterns,* vol. 4, no. 9, 100828, 2023, doi: https://doi.org/10.1016/j.patter.2023.100828.

[13] H. Javed, . S. El-Sappagh and . T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications," *Artificial Intelligence Review,* vol. 58, no. 12, 2025, doi: https://doi.org/10.1007/s10462-024-11005-9.

[14] O. Peretz, M. Koren and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Engineering Applications of Artificial Intelligence,* vol. 136, p. 108972, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108972.

[15] P. Pajila, B. Sheena, S. and A. G. , "A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications.," in *International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2023, doi: https://doi.org/10.1109/ICOSEC58147.2023.10276274.

[16] R. Shwartz-Ziv and A. Armoni, "Tabular Data: Deep Learning is Not All You Need," *ArXiv,* vol. 81, pp. 84-90, 2022, doi: https://doi.org/10.1016/j.inffus.2021.11.011.

[17] J. K. Sayyad, K. Attarde and N. Saadouli, "Optimizing e-Commerce Supply Chains With Categorical Boosting: A Predictive Modeling Framework," *IEEE Access,* vol. 12, pp. 134549-134567, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3447756.

[18] P. Mahesh and R. Soundrapandiyan , "Yield prediction for crops by gradient-based algorithms," *PLOS ONE,* vol. 19, no. 8, 2024, doi: https://doi.org/10.1371/journal.pone.0291928.

[19] D. Boldini, . F. Grisoni and S. A. Sieber, "Practical guidelines for the use of gradient boosting for molecular property prediction," *Journal of Cheminformatics,* vol. 15, p. 73, 2023, doi: https://doi.org/10.1186/s13321-023-00743-7.

[20] J. Singh, J. K. Sandhu and Y. Kumar, "Metaheuristic-based hyperparameter optimization for multi-disease detection and diagnosis in machine learning,"*Service Oriented Computing and Applications,* vol. 18, pp. 163-182, 2024, doi:10.1007/s11761-023-00382-8.

[21] J. A. Vásquez-Coronel, M. Mora and K. Vilch, "Review of multilayer extreme learning machine neural networks," *Artificial Intelligence Review,* vol. 56, p. 13691–13742, 2023, doi:10.1007/s10462-023-10478-4.

[22] D. Lusiyanti, S. Anam, W. F. Mahmudy and U. Sa'adah, "A Novel Hyperparameter Tuning Strategy Using Spider Monkey Optimization for Deep Neural Networks in Water Quality Prediction," Surabaya, Indonesia, 2025, doi: https://doi.org/10.1109/IES67184.2025.11160714.

[23] K. Taha, "Machine learning in biomedical and health big data: a comprehensive survey with empirical and experimental insights," *Journal of Big Data,* vol. 12, no. 61, pp. 1-39, 2025, doi:10.1186/s40537-025-01108-7.

[24] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbekel and B. V. Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *National Center for Biotechnology Information,* vol. 110, pp. 12-22, 2019.

[25] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg and O. Lyashevska, "Predictive analytics with gradient boosting in clinical medicine," *Big-data Clinical Trial Column,* vol. 7, no. 7, p. 152, 2019, doi: https://doi.org/10.1016/j.jclinepi.2019.02.004.

[26] R. Blanquero, E. Carrizosa, P. Ramirez-Cobo and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Computers & Operations Research,* vol. 135, p. 105456, 2021, doi: https://doi.org/10.1016/j.cor.2021.105456.

[27] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEEAccess,* vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[28] L. Muflikhah, W. W. F. Mahmudy and S. , "Improving Performance for Prediction of Hepatocellular Carcinoma Using Stacking Method of Support Vector Machine," *International Workshop on Innovations in Information and Communication Science and Technology* , 2020.

[29] S. Hanifi, A. Cammarono and H. Z.-B. , "Advanced hyperparameter optimization of deep learning models for wind power prediction," *Renewable Energy,* vol. 221, p. 119700, 2024, doi: https://doi.org/10.1016/j.renene.2023.119700.

[30] Y. Xu, . X. Zheng, Y. Li, X. Ye, H. Cheng and H. Wang, "Exploring patient medication adherence and data mining methods in clinical big data: A contemporary review," *Journal of Evidence-Based Medicine,* vol. 16, no. 3, pp. 342 - 375, 2023, doi: https://doi.org/10.1111/jebm.12548.

[31] M. R. Khurshid, S. Manzoor, T. Sadiq, L. Hussain and M. S. Khan, "Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction," *PLLOSO ONE,* vol. 6, 2025, doi: https://doi.org/10.1371/journal.pone.0310218.

[32] A. T. Tran, T. Zeevi and S. Payabvash, "Strategies to Improve the Robustness and Generalizability of Deep Learning Segmentation and Classification in Neuroimaging," *BioMedInformatics,* vol. 5, no. 2, p. 20, 2025, doi: https://doi.org/10.3390/biomedinformatics5020020.

[33] T.-T. Wong and P.-Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 32, no. 8, pp. 1586-1594, 2020, doi: https://doi.org/10.1109/TKDE.2019.2912815.

[34] A. Lei, Y. Mei, D. Ma, Z. Liu, w. Tao and F. Huang, "Two-stage transient stability assessment using ensemble learning and cost sensitivity," *Frontiers in Energy Research,* vol. 12, 2024, doi: https://doi.org/10.3389/fenrg.2024.1491846.

[35] I. Araf, A. Idri and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artificial Intelligence Review,* vol. 57, p. 80, 2024, doi: https://doi.org/10.1007/s10462-023-10652-8.