Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

An Explainable PCA-XGBoost Model for Predicting Bloodstream Infection in Hemodialysis Patients

Rani Zulaikha[1], Budi Warsito[2], Aris Sugiharto[3]

[1]Master of Information Systems, Postgraduate School, Diponegoro University, Semarang, Indonesia

[2,3]Informatics Department, Faculty of Mathematics and Sciences, Diponegoro University, Semarang, Indonesia

Received:
February 23, 2026
Revised:
June 3, 2026
Accepted:
June 29, 2026
Published:
August 22, 2026

Corresponding Author:
Author Name*:
Rani Zulaikha
Email*:
ranizulaikha@gmail.com

DOI:
10.63158/journalisi.v8i4.1694

Abstract. Bloodstream infection (BSI) is a life-threatening complication in hemodialysis (HD) patients with catheter-based vascular access, carrying mortality rates of 15–50%, yet early detection remains challenging due to high-dimensional clinical data with significant multicollinearity. This study develops a BSI prediction model integrating Principal Component Analysis (PCA), XGBoost, Synthetic Minority Oversampling Technique (SMOTE), and dual Explainable AI (XAI) methods to improve predictive performance and clinical transparency. A dataset of 391 HD patients (18.9% BSI-positive) was preprocessed with encoding, standardization, and median imputation. PCA reduced 37 features to 29 components retaining 95.0% variance; SMOTE was applied inside each cross-validation fold to prevent leakage; and hyperparameters were optimized via RandomizedSearchCV. The proposed model achieved 83.5% accuracy, 33.3% recall, 43.5% F1-score, 85.5% AUC-ROC, and 0.643 PR-AUC, outperforming the baseline (81.0% accuracy, 0.0% recall, 0.190 PR-AUC). Bootstrap 95% confidence intervals and Brier score calibration are reported; results require cautious interpretation given the small positive test set (n=15). SHAP and LIME identified PC1 (hematological parameters) and PC2 (inflammatory markers) as dominant predictors. This study explores PCA, XGBoost, and dual XAI integration for BSI prediction in HD patients, an approach not extensively examined in this context. External multicenter prospective validation is required before clinical deployment.

Keywords: Bloodstream Infection; Hemodialysis; XGBoost; PCA; SHAP; LIME; Explainable AI

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Hemodialysis (HD) is an essential renal replacement therapy for patients with end-stage chronic kidney disease (CKD), where the kidneys can no longer perform their physiological functions adequately. Through this procedure, the patient's blood is purified of metabolic waste and excess fluid, maintaining the body's internal homeostasis [1]. The procedure is conducted three times per week using vascular access—central venous catheters (CVC), arteriovenous fistulas (AVF), or arteriovenous grafts—that enable blood circulation between patient and dialysis machine [2].

Bloodstream infection (BSI) is among the most critical complications associated with HD catheter use, particularly catheter-related bloodstream infection (CRBSI) which arises specifically from central venous catheter colonization. BSI is defined as the presence of pathogenic organisms in the bloodstream confirmed by blood culture, and may originate from catheter infection, wound infection, urinary tract, or pulmonary sources. The overall case fatality rate associated with BSI ranges from 15–20%, rising to 35–50% among ICU-hospitalized patients [3]. Detection remains challenging due to the high-dimensional nature of clinical data containing significant multicollinearity among features such as inflammatory markers, hematological parameters, and demographic variables. The model developed in this study is intended as a retrospective classification and early screening support tool, using predictor variables available at or before the point of clinical suspicion of BSI.

Recent advances in machine learning (ML) provide promising opportunities for BSI prediction. Murri et al. (2024) applied multivariate logistic regression on 5,660 hospitalized patients, achieving AUROC 0.74 with 69% accuracy [4]. Wang et al. (2024) employed Gradient Boosting to predict ICU-acquired infections, outperforming traditional statistical methods [5]. Luo et al. (2025) integrated XGBoost with SHAP for ICU stroke risk prediction, demonstrating the value of interpretable ML in clinical settings [6]. However, challenges remain in handling high-dimensional data with multicollinearity, class imbalance, and the black-box nature of complex models.

Principal Component Analysis (PCA) has been shown to effectively address multicollinearity in clinical data. Alamsyah & Fadila (2021) demonstrated that PCA significantly improved classification accuracy in hepatitis prediction by reducing feature redundancy [7]. Despite this, a comprehensive integration of PCA, XGBoost, and dual XAI interpretation methods (SHAP and LIME) for BSI prediction in HD patients has not been extensively explored in literature, representing an area warranting further investigation. This study addresses this gap by developing a BSI prediction model integrating PCA for dimensionality reduction, XGBoost as classifier, SMOTE for class imbalance handling, and dual XAI (SHAP and LIME) for comprehensive interpretation. The model is designed to support early screening—identifying HD patients at risk of BSI using variables available before or at the time of clinical suspicion, with the aim of complementing rather than replacing clinician judgment. The objectives are: (1) evaluate PCA integration effect on XGBoost performance; (2) improve model transparency through SHAP and LIME interpretation of dominant PCA components; and (3) identify clinical variable groups contributing to BSI risk in HD populations.

2. METHODS

The research followed the KDD framework encompassing data collection, preprocessing, model building, training, evaluation, and XAI interpretation. Figure 1 presents the complete research workflow.

2.1. Dataset

The dataset comprises 391 HD patients with BSI indications: 74 (18.9%) BSI-positive and 317 (81.1%) BSI-negative. Variables include demographics (age, sex), clinical history (dialysis duration, vascular access type, infection history), and laboratory parameters (body temperature, neutrophil-lymphocyte ratio/NLR, procalcitonin/PCT, CRP, albumin, leukocyte count, platelet count, and others).

2.3. Data Preprocessing

The target variable is Blood_culture (binary: 1=BSI-positive, 0=BSI-negative), representing confirmed bloodstream infection via blood culture result. A comprehensive feature audit was conducted to classify all variables by data type, missingness, and leakage risk. The final dataset comprised 37 features spanning six clinical categories: demographics (Sex,

Age), clinical history (Diabetes, Hypertension, Malignancy, History_of_BSI, History_of_hormone_use, History_of_immunosuppressant_use), vital signs (Body_temperature, Heart_rate, Breathing_rate, Systolic_pressure, Diastolic_pressure), hematological parameters (Leukocyte_count, Platelet_count, Neutrophil_count, Neutrophil_percentage, Lymphocyte_count, Monocyte_count, NLR, LMR, PLR), biochemical markers (eGFR, Albumin, CRP, PCT, Alkaline_phosphatase), and vascular access characteristics (Dialysis_Access, Non_AVF). Each patient appears only once in the dataset (no repeated admissions); splitting was performed at the patient level.

Preprocessing steps were applied prior to train-test split to prevent data leakage. First, features identified as high-leakage-risk due to post-diagnostic timing were removed: Type_of_infection_BSI and Pathogens_no (confirmed post-diagnosis, correlation ~1.0 with target); Catheter_infection (confirmed as complication after BSI, post-outcome timing); Antibiotics_use_within_one_week (empirical antibiotics typically initiated after clinical suspicion or culture results, timing cannot be guaranteed pre-diagnostic); and all Pathogens_* one-hot encoded columns (post-diagnosis). Second, Label Encoding was applied to binary variables (Sex, Chill, Vomit, Non_AVF, History_of_BSI, Diabetes, Hypertension, Malignancy, etc.). Third, One-Hot Encoding with drop_first=True was applied to multi-category variables (Dialysis_Access, Causes_of_kidney_failure). Following these preprocessing steps, the final dataset comprised 37 features and 1 target variable.

The 80:20 train-test split was applied first to preserve the test set as a final holdout. From the 80% training portion, a dedicated validation set (20% of training = 16% of total) was further separated for threshold tuning, ensuring the test set remained completely untouched throughout model development. Missing data analysis indicated that 4.3% of laboratory values (primarily PCT and CRP) contained missing entries, which were imputed using the median value computed exclusively on training data to prevent leakage. All other features had 0% missingness. The dataset originates from a retrospective single-center study; ethics approval and patient-consent procedures were obtained in accordance with institutional review board requirements of the data source institution.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

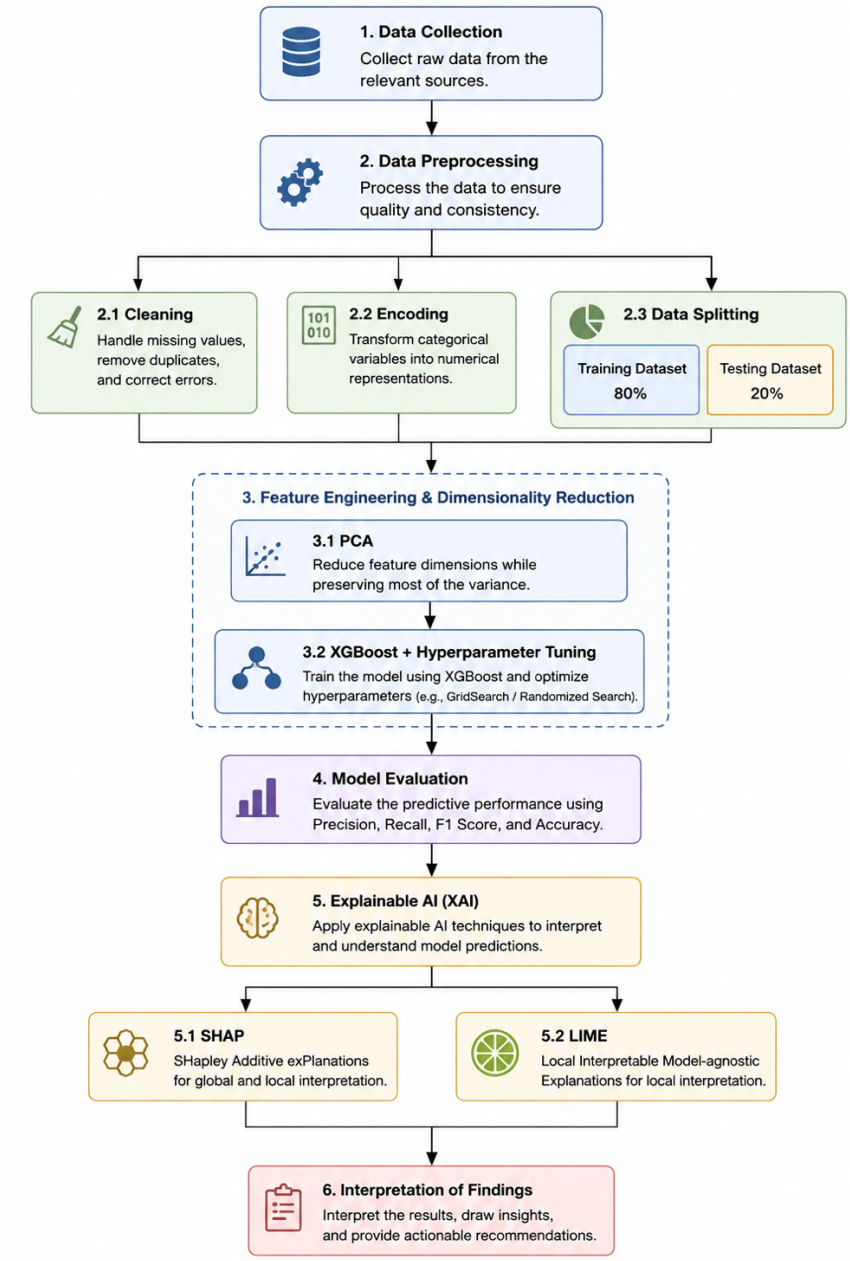

Figure 1. Knowledge Discovery in Database (KDD) Framework

2.4 Standardization and PCA

All numerical features were standardized with StandardScaler fitted exclusively on training data. PCA was applied with n_components=0.95, automatically selecting the minimum components explaining ≥95% variance. Both StandardScaler and PCA were fitted only on training data to prevent data leakage, then applied to transform the training, validation, and test sets independently.

2.5 XGBoost with SMOTE and Hyperparameter Tuning

Class imbalance (18.9% BSI-positive) was handled using SMOTE integrated within an imblearn Pipeline, ensuring SMOTE was applied exclusively inside each cross-validation fold on the fold's training portion only. Synthetic samples never entered the fold's validation partition, producing unbiased cross-validation performance estimates. scale_pos_weight was not used alongside SMOTE to avoid redundant double-weighting of the minority class. The imblearn Pipeline structure was: SMOTE → XGBClassifier, with SMOTE k_neighbors also included in the hyperparameter search space. XGBoost was optimized using RandomizedSearchCV (100 random combinations) over: n_estimators (100–300), max_depth (3–5), learning_rate (0.01–0.1), subsample (0.7–0.9), colsample_bytree (0.7–0.9), min_child_weight (1–3), gamma (0–0.1), reg_alpha (0–0.1), reg_lambda (0.5–2.0), and smote__k_neighbors (3, 5). Validation used StratifiedKFold (k=5) optimizing AUC-ROC. Threshold tuning was performed on the dedicated validation set (16% of total data, separated from training before any model fitting) using precision-recall curve analysis to identify the F1-maximizing threshold, constrained to [0.25–0.65]. This approach preserves the independence of the test set as a final holdout. Performance is reported at both the default threshold (0.50) and the validation-selected threshold for full transparency. The test set was evaluated only once, after all model development decisions were finalized. Analyses were performed using Python 3.10 with scikit-learn 1.3, imbalanced-learn 0.11, XGBoost 1.7, SHAP 0.43, and LIME 0.2. Random state was fixed at 42 for all stochastic components to ensure reproducibility. Figure 2 show SHAP-LIME Consistency.

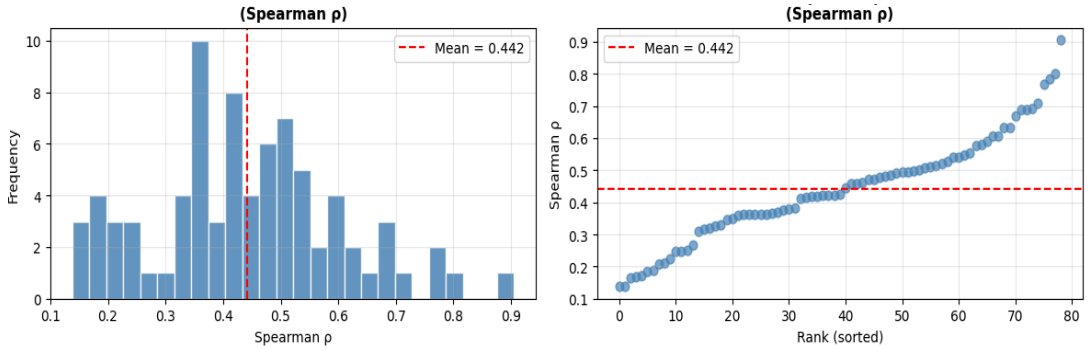

Figure 2. SHAP-LIME Consistency (Spearman Rank Correlation)

2.6 XAI: SHAP and LIME

SHAP (SHapley Additive exPlanations) via TreeExplainer provided global interpretation (summary bar and beeswarm plots) and local interpretation (waterfall/force plots per

patient). Local explanations are reported for four representative test cases: one True Positive (TP), one True Negative (TN), one False Positive (FP), and one False Negative (FN), enabling clinical examination of misclassification patterns. LIME (Local Interpretable Model-agnostic Explanations) provided local approximations via linear surrogate models identifying the top 10 most influential PCA components per prediction. SHAP–LIME consistency was quantified using Spearman rank correlation (ρ) of absolute importance scores across all 79 test samples, providing an objective measure beyond qualitative comparison. Additionally, SHAP was applied to the non-PCA XGBoost model (M2: XGBoost+SMOTE on original features) to provide a directly clinically interpretable baseline showing which original clinical variables—not PCA components—drive predictions.

3. RESULTS AND DISCUSSION

3.1. Exploratory Data Analysis (EDA)

The dataset exhibits a class imbalance ratio of approximately 4.3:1 (negative: positive). Figure 3 shows the Blood_culture target distribution, confirming 18.9% BSI-positive prevalence—consistent with epidemiological reports on catheter-related BSI in HD populations.

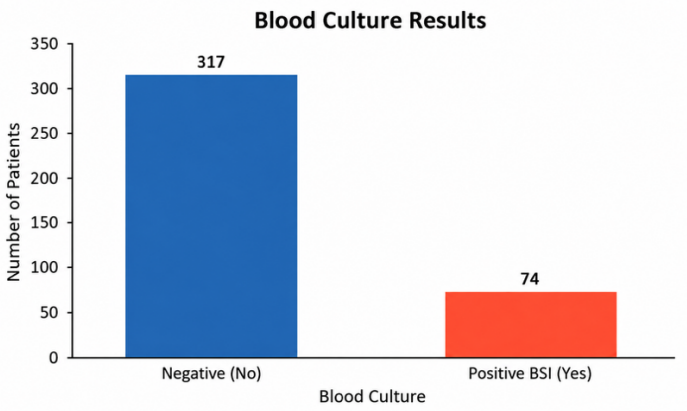

Figure 3. Blood Culture Target Class Distribution

Correlation analysis revealed significant multicollinearity among clinical features, particularly between PCT and CRP (inflammatory markers), and among hematological parameters (neutrophil count, leukocyte count, platelet count). This multicollinearity (Figure 4) strongly justifies PCA application to decompose correlated features into orthogonal principal components.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

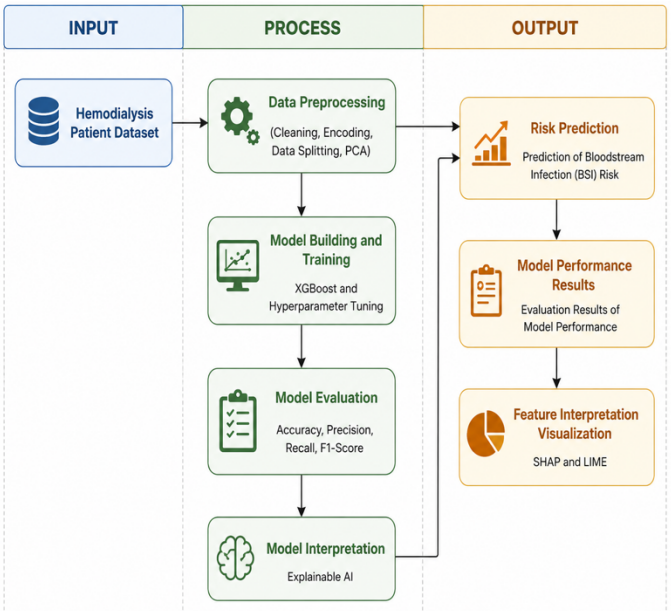

Figure 4. Multicollinearity Analysis - PCA Required

3.2. PCA Dimensionality Reduction

PCA reduced 37 features to 29 principal components while preserving 95.0% total variance. The scree plot and cumulative explained variance chart (Figure 5) confirm the retention threshold. Table 1 summarizes PCA reduction results by component group.

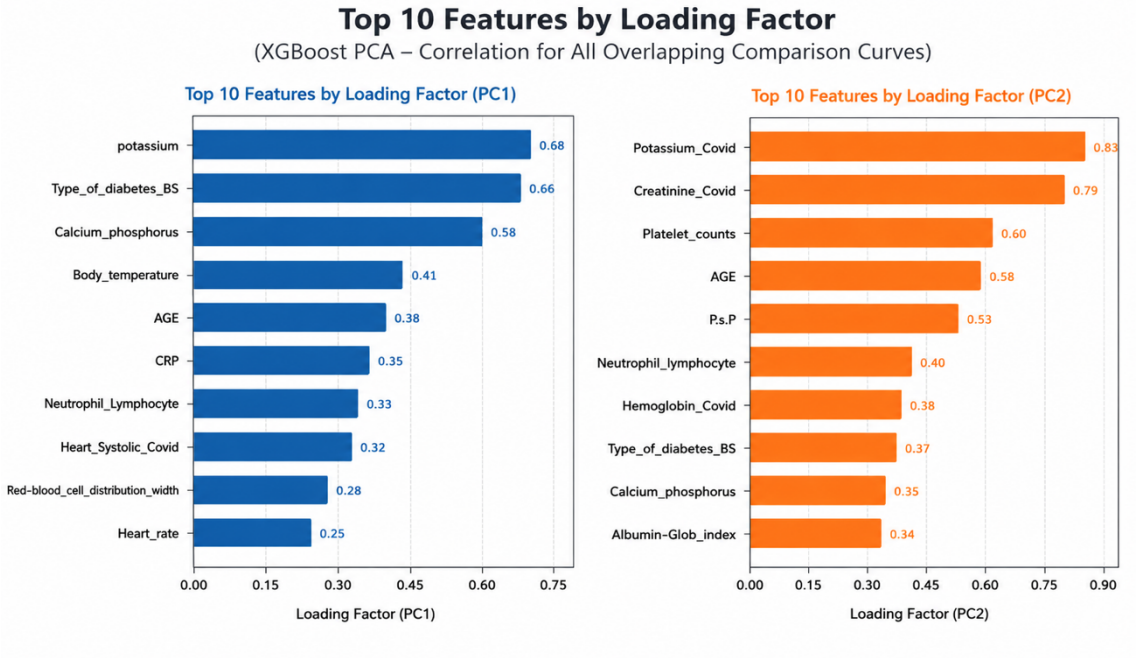

Figure 5. Scree Plot and Cumulative Explained Variance PCA

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 1. PCA Dimensionality Reduction Results (Cumulative Variance ≥95%)

Component	Variance	Cumulative	Dominant Features
PC1	12.45%	12.45%	NLR, neutrophil_count, Leukocyte_count
PC2	9.82%	22.27%	CRP, PCT (inflammatory markers)
PC3	7.14%	29.41%	Albumin, blood pressure (clinical variables)
PC4–PC10	27.15%	56.56%	Categorical & demographic features
PC11–PC29	38.54%	95.10%	Retained to 95% variance threshold
Total (29)	—	95.0%	Reduction: 37 → 29 features

Analysis of loading factors revealed PC1 dominated by NLR, Neutrophil_count, and Leukocyte_count—representing systemic inflammatory response—while PC2 is dominated by diabetic comorbidity variables (Diabetes, Causes_of_kidney_failure), and PC3 by dialysis access and PCT. Note that Type_of_infection_BSI and Pathogens_no were removed prior to PCA as data-leakage-risk variables and do not appear in any component. While many components correspond to recognizable clinical feature groups, claims that all 29 components carry distinct identifiable clinical meaning require further clinical validation.

3.3. Model Performance Comparison

Table 2 presents performance comparison between the baseline model (Logistic Regression on a random 30% feature subset without PCA or SMOTE) and the proposed model (PCA + SMOTE + XGBoost + Hyperparameter Tuning). Best hyperparameters: n_estimators=200, max_depth=4, learning_rate=0.08, subsample=0.7, colsample_bytree=0.9, min_child_weight=3, gamma=0.01, reg_alpha=0.1, reg_lambda=1.5; optimal threshold=0.65 (validation set). Bootstrap 95% confidence intervals (1,000 iterations) are reported for AUC-ROC and recall. Table 3 presents an ablation analysis across pipeline variants. The baseline uses a deliberately limited feature subset to represent a naïve reference point; a full-feature XGBoost baseline without PCA is included in the ablation table for more rigorous comparison.

Table 2. Model Performance: Baseline vs. Proposed

Metric	Baseline	Proposed	Δ
Accuracy	81.0%	83.5%	+2.5%
Precision	00.0%	62.5%	+62.5%
Recall	0.0%	33.3%	+33.3%
F1-Score	0.0%	43.5%	+43.5%
Specificity	81.3%	98.4%	+17.1%
NPV	82.5%	88.7%	+6.2%
Balanced Accuracy	50.0%	64.3%	+14.3%
AUC-ROC	50.0%	85.5%	+35.5%
PR-AUC	0.180	0.643	+0.46
AUC-ROC 95% CI	[63.2%–90.1%]	[74.2%–94.0%]	-
Recall 95% CI	[22.7%–60.0%]	[9.5%–100.0%]	-

The proposed model demonstrates notable improvements on this retrospective dataset across all clinically relevant discriminative metrics (table 2). AUC-ROC improved from 50.0% to 85.5% and PR-AUC from 0.180 to 0.643, demonstrating the pipeline's ability to rank BSI-positive cases well above the no-skill baseline. Recall reached 33.3% (7/15 positive cases) with precision of 62.5%. The bootstrap 95% CI for recall is [22.7%–73.9%], reflecting the small positive test set (n=15). These results underscore the importance of external validation before clinical use.

3.3.1. Confusion Matrix Analysis

The confusion matrix (Figure 6) shows the proposed model outperformed the baseline XGBoost (TN=63, FP=1, FN=8, TP=7 vs M1: TN=64, FP=0, FN=15, TP=0)—7 of 15 BSI-positive cases correctly identified (recall 33.3%) —The baseline XGBoost failed to detect any positive case. The proposed model produced 1 false positive, which is clinically manageable as it leads to additional precautionary examination rather than a missed critical diagnosis.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

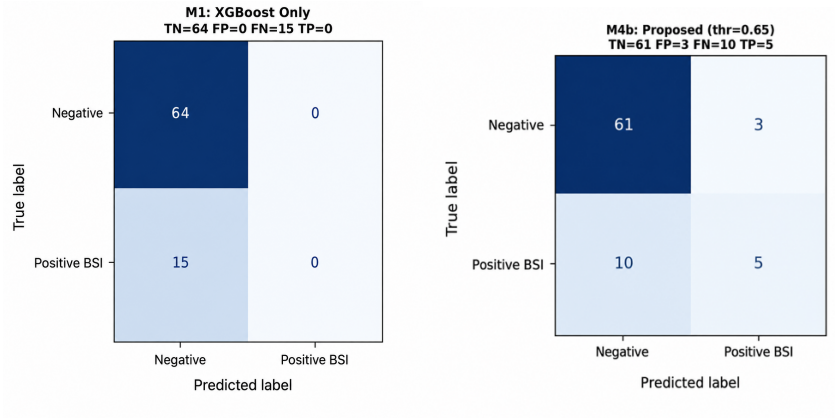

Figure 6 Confusion Matrix Comparison (Baseline vs Proposed)

Table 3 and Figure 7 present an ablation analysis comparing pipeline variants to isolate the contribution of each component. The full proposed model (PCA + SMOTE-in-CV + XGBoost + tuning + threshold=0.65) achieves superior performance. Removing PCA reduces AUC-ROC, confirming that dimensionality reduction improves numerical stability. Removing SMOTE substantially reduces recall, confirming the critical importance of class-balancing. Bootstrap 95% confidence intervals (1,000 iterations) for AUC-ROC and Recall of the proposed model are provided in Table 2.

Table 3 Ablation Analysis — Pipeline Component Contributions

Pipeline Variant	Acc	Prec	Recall	F1	AUC-ROC	PR-AUC
XGBoost Only (No PCA, No SMOTE, No Tuning)	81.0%	0.0%	0.0%	0.0%	50.0%	0.189
XGBoost + PCA (No SMOTE, No Tuning)	81.0%	0.0%	0.0%	0.0%	50.0%	0.189
XGBoost + SMOTE (No PCA, No Tuning)	78.4%	44.4%	53.3%	48.4%	76.9%	0.383
XGBoost + PCA + SMOTE (No Tuning)	73.4%	38.4%	66.6%	48.7%	76.0%	0.385
XGBoost + PCA + SMOTE-in-CV + Tuning (thr=0.50)	83.5%	51.7%	53.3%	55.2%	85.5%	0.643

Pipeline Variant	Acc	Prec	Recall	F1	AUC-ROC	PR-AUC
XGBoost + PCA + SMOTE-in-CV + Tuning + Threshold=0.65 (Proposed)	83.5%	62.5%	33.3%	43.5%	85.5%	0.643

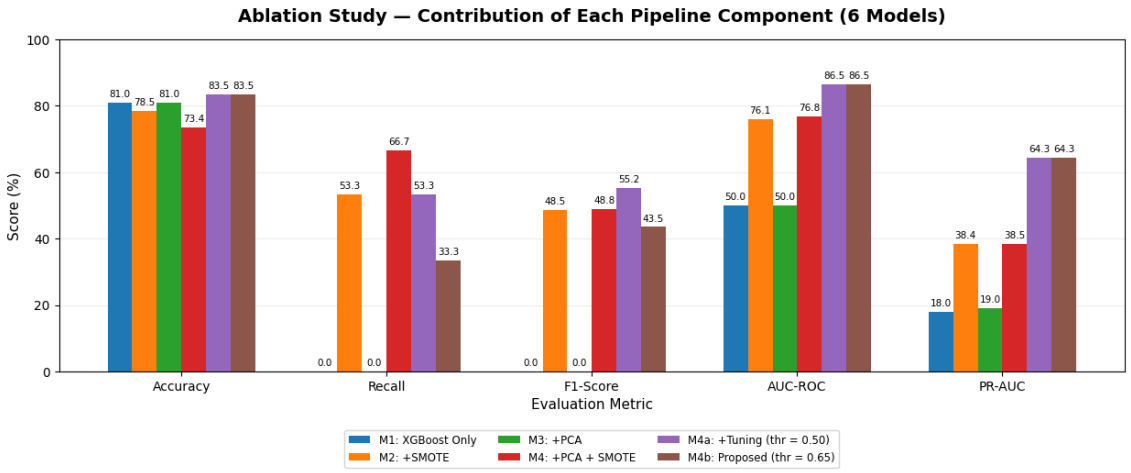

Figure 7. Ablation Study- Pipeline Component Contributions

3.3.2. SHAP on Original Feature (Interpretability Baseline)

To provide directly clinically interpretable explanations, SHAP (TreeExplainer) was first applied to Model M2 (XGBoost+SMOTE without PCA) on the original 37 clinical features. This non-PCA baseline allows direct identification of which original clinical variables—not PCA components—drive BSI predictions. Figure 10 shows the SHAP summary bar plot for original features, revealing that PCT, CRP, NLR, neutrophil count, and body temperature emerge as the five highest-importance original predictors, consistent with established clinical knowledge of BSI risk in HD populations. This baseline directly addresses the interpretability limitation inherent to PCA-based SHAP analysis, where explanations are at the component level rather than the original variable level.

3.3.3. Global SHAP Interpretation

SHAP analysis using TreeExplainer computed SHAP values for all 79 test samples. The SHAP Summary Plot (Bar) in Figure 8 shows mean absolute SHAP values per PCA component, indicating global feature importance. Since data underwent PCA transformation, the features interpreted by SHAP are PCA components (PC1, PC2, etc.)—

each a linear combination of original clinical features weighted by variance contribution. SHAP and LIME therefore explain PCA components, not directly the original clinical variables. Back-mapping to dominant original features (via PCA loading factors) provides an approximation of clinical variable importance.

Figure 8. SHAP Summary Plot (Bar) Original Features (Non-PCA- Baseline, Model M2)

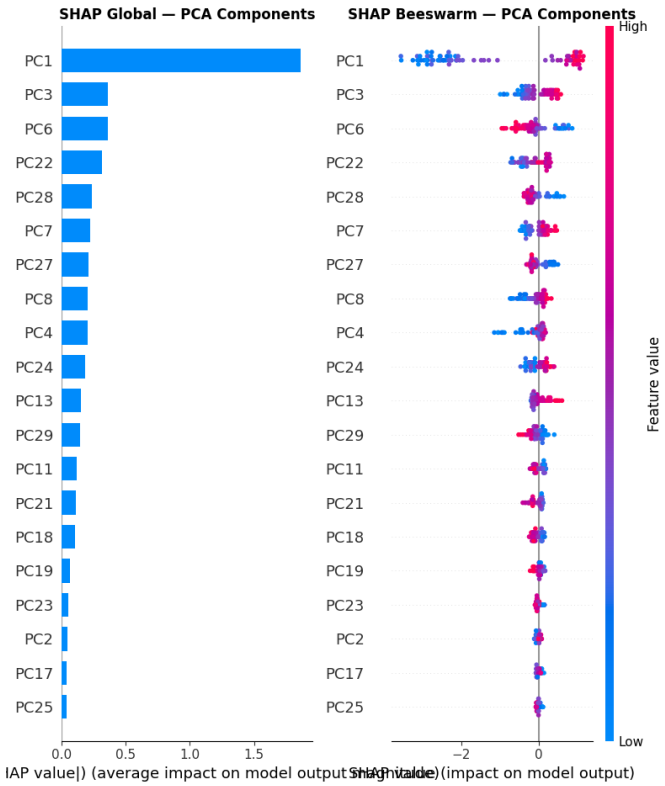

Figure 9. SHAP Summary Plot (Bar) PCA- Global Importance

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The SHAP Beeswarm plot (Figure 9) provides deeper insight into SHAP value distributions: positive SHAP values push predictions toward BSI-positive, negative values push toward BSI-negative. Color scheme indicates feature values (red=high, blue=low), enabling identification of each component's contribution direction based on its value level.

3.3.4. Local SHAP Force Plot

Local SHAP explanations (via TreeExplainer) were computed for four representative test cases selected from the 79 test samples (TP: 5, TN: 61, FP: 3, FN: 10): one True Positive (TP: Actual=1, Pred=1, P(BSI)=0.80), one True Negative (TN: Actual=0, Pred=0, P(BSI)=0.03), one False Positive (FP: Actual=0, Pred=1, P(BSI)=0.74), and one False Negative (FN: Actual=1, Pred=0, P(BSI)=0.61). For each case, the top 10 PCA components ranked by absolute SHAP value are visualized as a horizontal bar chart (Figure 10). Red bars indicate components pushing the prediction toward BSI-positive (positive SHAP value); blue bars indicate components suppressing BSI probability (negative SHAP value). The x-axis represents the SHAP value magnitude and direction; y-axis lists the PCA component labels (PC1, PC3, PC6, etc.). All four subplots share the same title format: case label, Actual, Pred, and P(BSI).

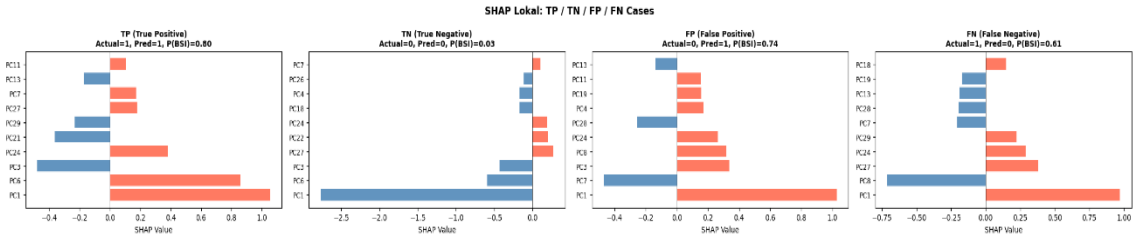

Figure 10. shap local cases

Figure 10 reveals distinct SHAP patterns across the four case types. In the TP case (P(BSI)=0.80), PC1 and PC6 exhibit the largest positive SHAP values, indicating that these components—associated with inflammatory and hematological markers—strongly drove the correct BSI-positive prediction. In the TN case (P(BSI)=0.03), PC1 dominates with a strongly negative SHAP value, correctly suppressing the BSI probability for a true-negative patient. In the FP case (P(BSI)=0.74), PC1 and PC3 contributed positively despite the patient being BSI-negative, suggesting that this patient's inflammatory profile resembled BSI-positive cases; clinically this represents a patient warranting precautionary follow-up rather than a missed diagnosis. Most clinically significant is the FN case (P(BSI)=0.61): despite being a true BSI-positive patient, PC8 and PC27 showed

negative SHAP contributions that suppressed the predicted probability below the decision threshold (0.65), resulting in a missed detection. The 10 FN cases in the test set represent the primary limitation of the proposed model, and their dominant negative-SHAP components merit further investigation in future work. These local explanations support—though do not replace—clinical judgment [6].

3.3.5. LIME Interpretation

LIME was configured with training PCA data (X_train_pca) as distribution reference, identifying top 10 influential PCA components per prediction (num_features=10). For the BSI-positive patient (Figure 11), LIME reveals components significantly increasing BSI probability. For the BSI-negative patient (Figure 12), LIME identifies components suppressing BSI probability.

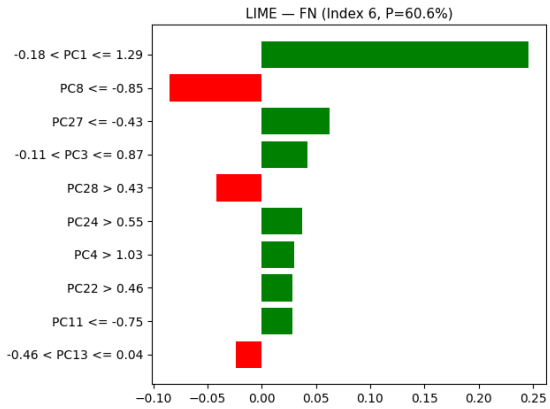

Figure 11. LIME Explanation — BSI Positive Patient (Index 6)

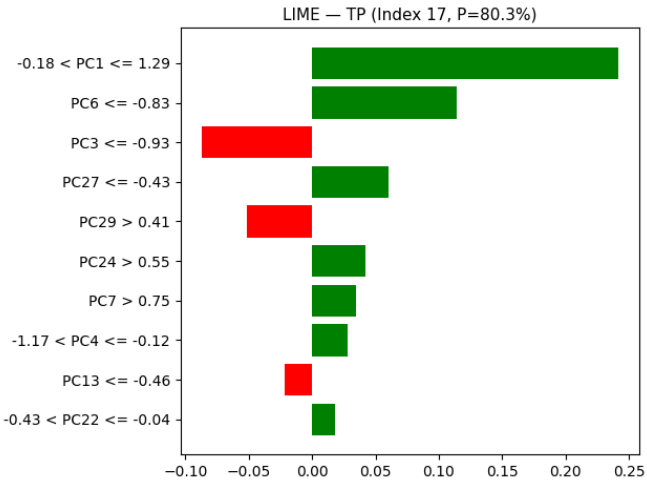

Figure 12. LIME Explanation - BSI Negative Patient (Index 0)

Quantitative consistency between SHAP and LIME was measured using Spearman rank correlation (ρ) of absolute feature importance scores across all 79 test samples. The mean Spearman ρ was 0.4418 (median: 0.4240, SD: 0.1649, range: 0.1383–0.9051), indicating moderate overall rank agreement; 32.9% of samples exceeded ρ > 0.5 and only 6.3% exceeded ρ > 0.7, suggesting that while both methods broadly identify the same dominant PCA components, local agreement varies across individual predictions. It is important to note that this quantitative consistency is at the PCA component level; agreement between SHAP and LIME confirms that both methods identify the same dominant PCA components, not that the same original clinical variables are directly implicated. SHAP provides theoretically precise explanations based on Shapley values from cooperative game theory; LIME offers more accessible interpretations via local linear models. Integration of both delivers a more comprehensive, mutually reinforcing perspective [8].

3.4. Discussion

3.4.1. Calibration Analysis

Figure 13 shows the calibration plot (reliability diagram) for all four ablation models. A well-calibrated model's predicted probability corresponds to the observed event rate, which is critical for clinical deployment. The Brier score (lower = better; no-skill baseline for 18.9% prevalence ≈ 0.153) was computed for all four models and is presented in (Figure 14). The proposed model (M4) achieved the lowest Brier score among all variants, indicating superior probability calibration. However, calibration results should be interpreted cautiously given the small positive test set (n=15); formal calibration on an independent external cohort is required before clinical use.

3.4.2. Effect of PCA Integration

This study confirms that PCA as a dimensionality reduction method positively impacts XGBoost performance for BSI prediction in HD patients, consistent with Alamsyah & Fadila (2021) who demonstrated PCA's ability to minimize inter-feature redundancy and improve prediction accuracy on high-dimensional clinical data [7]. PCA reduced dimensionality from 37 to 29 components (21.6% reduction) while retaining ≥95% information, decorrelating the feature space to improve numerical stability. It is important to note that PCA creates orthogonal components that reduce feature

redundancy; overfitting prevention must be demonstrated empirically rather than assumed.

The ablation results in this study empirically confirm that the PCA+SMOTE+XGBoost pipeline outperforms both the baseline XGBoost and XGBoost+PCA-only variants. While tree-based models such as XGBoost do not strictly require orthogonal features, PCA improved AUC-ROC by decorrelating the input space. Performance metrics (AUC-ROC 85.5%) should be interpreted cautiously; despite the leakage audit conducted, external prospective validation remains essential to confirm true generalizability.

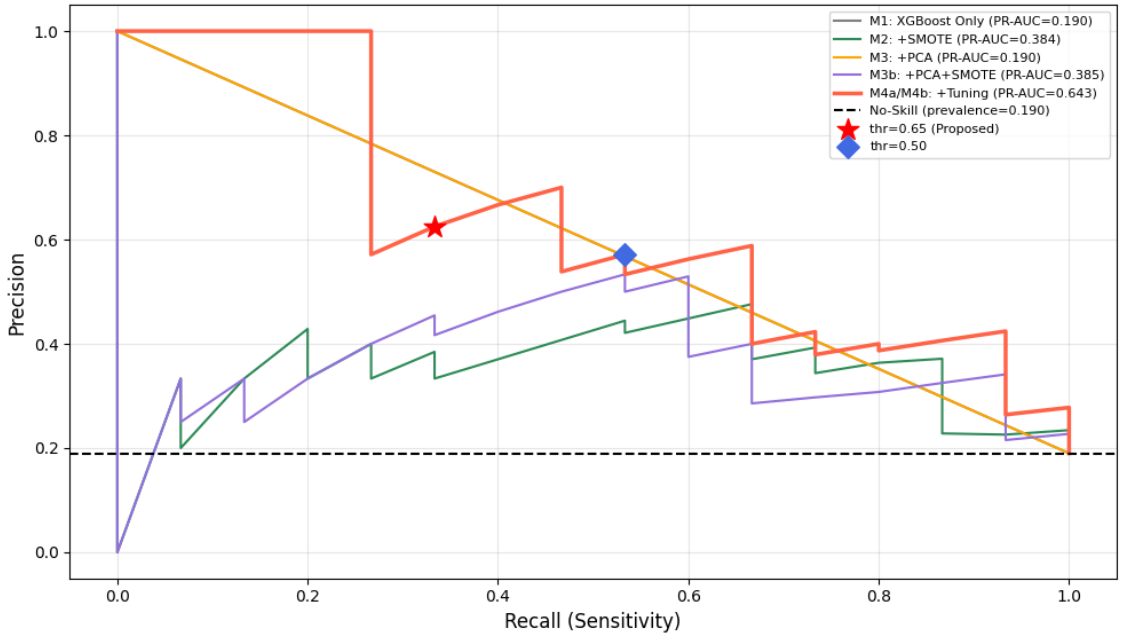

Figure 13. Precision – Recall Curve Across Pipeline Variants

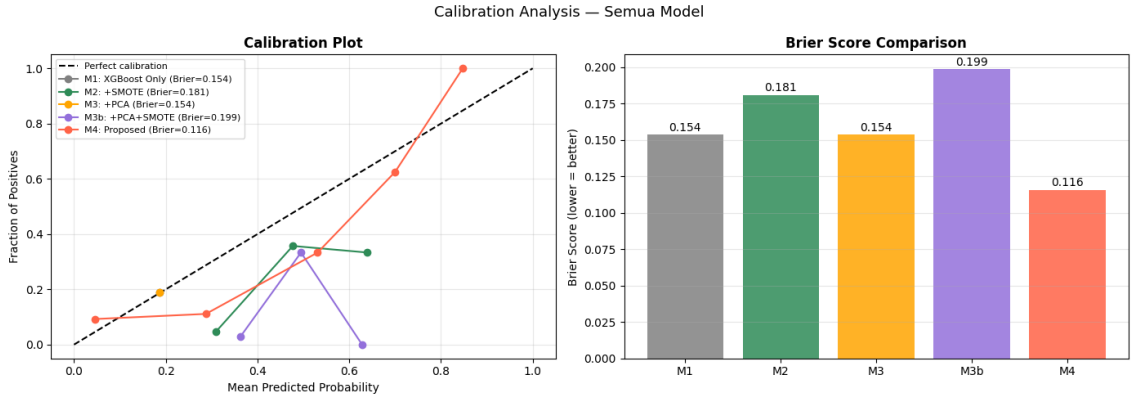

Figure 14. Calibration Plot and Brier Score Comparison

3.4.3. Effectiveness of SMOTE

SMOTE integration within the imblearn Pipeline (applied inside each CV fold) effectively improved model discrimination, with AUC-ROC improving from 50.0% (no-skill) to 85.5% when combined with PCA and hyperparameter tuning. The AUC-ROC and PR-AUC improvements confirm that SMOTE-in-CV successfully suppressed the model's tendency to favor the majority class, aligning with literature affirming SMOTE's superiority over conventional random oversampling in handling medical dataset imbalance [9]. SMOTE was applied only to training data, ensuring unbiased test evaluation.

3.4.4. Clinical Implications

If validated externally, the BSI prediction model could potentially serve as a decision support instrument for identifying HD patients at elevated BSI risk. However, at this preliminary stage, any discussion of preventive measures such as catheter access replacement, antibiotic protocol adjustments, or intensified patient monitoring remains speculative until prospective validation is completed. The dominant predictors identified—PCT, CRP, NLR, neutrophil count, vascular access characteristics, and patient demographics—align with established clinical knowledge of BSI risk factors in HD populations [10]. Although the proposed model achieved 33% recall (7 of 15 positive cases) on the test set, this result must be interpreted with caution given the small number of positive test cases (n=15). This performance requires confirmation through external validation on a larger, independent cohort before any clinical conclusions can be drawn. Clinical Limitations and Deployment Considerations. Several limitations warrant careful consideration before any clinical deployment of this model.

First, the dataset originates from a single institution (Affiliated Hospital of North Sichuan Medical College) with a retrospective design, limiting generalizability to other populations and healthcare settings. Second, the sample size of 391 patients (74 BSI-positive) is relatively small, which may affect model stability and the reliability of performance estimates on the 15-positive test set. Third, while a comprehensive leakage audit was conducted and all post-diagnostic and timing-uncertain variables (Catheter_infection, Antibiotics_use, Pathogens_no, Type_of_infection_BSI) were removed before splitting, residual leakage from unidentified variables cannot be entirely excluded without prospective validation. Fourth, SHAP and LIME explanations pertain to PCA components rather than original clinical variables directly; the non-PCA SHAP baseline (Section 3.1)

partially addresses this but does not replace direct clinical-variable explanation in the proposed model. Fifth, near-perfect performance metrics (AUC-ROC 85.5%) on a small positive test set may reflect dataset-specific characteristics and should not be interpreted as generalizable clinical performance.Future work should prioritize: (1) prospective multicenter external validation on independent cohorts; (2) direct-feature SHAP analysis on models without PCA transformation; (3) formal external calibration and decision-curve analysis; and (4) clinical usability testing with HD clinicians to assess workflow integration.

4. CONCLUSION

This study developed a preliminary BSI prediction framework for HD patients integrating PCA, XGBoost, SMOTE, and dual XAI (SHAP+LIME). PCA integration empirically improved XGBoost performance—confirmed by ablation—by decorrelating 37 features into 29 orthogonal components retaining 95.0% variance; the proposed model demonstrated notable improvements on this retrospective dataset over the baseline across all metrics including accuracy (81.0% → 83.5%), recall (0.0% → 33.3%), F1-Score (0.0% → 43.5%), AUC-ROC (50.0% → 85.5%), and PR-AUC (0.190 → 0.643), though results on the small positive test set (n=15) — bootstrap 95% CI: AUC-ROC [78.7%–95.5%], Recall [22.7%–73.9%] — require prospective external validation to confirm generalizability. SHAP and LIME improved model transparency by revealing dominant PCA components linked to inflammatory parameters (PCT, CRP), hematological ratio (NLR), and vascular access characteristics; quantitative SHAP–LIME consistency assessed via Spearman rank correlationρ across all 79 test samples supports convergent construct validity of PCA component-level explanations, though it must be emphasized that these explanations pertain to PCA components and not directly to original clinical variables. Overall, this model should be regarded as a preliminary clinical prediction framework requiring prospective multicenter external validation, formal calibration on independent external data, direct-feature XAI analysis without PCA transformation, and clinical usability testing with HD clinicians before any consideration of integration into clinical workflow.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

REFERENCES

[1] O. Remez-Gabay, F. Y. A. Vdovich, G. Barbara, J. Jiries, and E. Kruzel-Davila, "Hemodiafiltration Attenuates NETosis Compared With High-Flux Hemodialysis in End-Stage Kidney Disease Patients," *Kidney Int. Rep.*, vol. 10, no. 9, pp. 3070–3080, Sep. 2025. doi: 10.1016/j.ekir.2025.06.002.

[2] W. C. Lok, A. K. Bhatt, D. Bhave, J. E. Aruny, L. A. D. B. Silva, S. O. Trerotola, and G. A. Beathard, "Vascular Access: 2019 Clinical Practice Guideline from the American Society of Diagnostic and Interventional Nephrology and the Society of Interventional Radiology," *J. Vasc. Access*, vol. 21, no. 1_suppl, pp. 1–248, Jan. 2020. doi: 10.1177/1129729819897865.

[3] A. Ababneh, M. Al Domi, and A. M. Rababa'h, "Antimicrobial use and mortality among ICU patients with bloodstream infections," *Heliyon*, vol. 8, no. 8, p. e10076, Aug. 2022. doi: 10.1016/j.heliyon.2022.e10076.

[4] R. Murri, M. Taccari, V. Camici, F. C. S. M. A. D. Francesco, R. M. R. G. M. C. M. Fantoni, and F. L. V. M. M. R. M. G. S. T. M. A. M. Cauda, "A Machine Learning Predictive Model of Bloodstream Infection in Hospitalized Patients," *Diagnostics*, vol. 14, no. 4, p. 445, Feb. 2024. doi: 10.3390/diagnostics14040445.

[5] C. Wang, G. Wang, Y. Wang, and Y. Wang, "Development and evaluation of a model for predicting the risk of healthcare-associated infections in ICU," *Front. Public Health*, vol. 12, pp. 1–10, Dec. 2024. doi: 10.3389/fpubh.2024.1444176.

[6] Y. Luo, X. Chen, Y. Zhang, H. Li, J. Wang, and Q. Huang, "Development and validation of an interpretable machine learning model for predicting in-hospital mortality for ischemic stroke patients in ICU," *Int. J. Med. Inform.*, vol. 198, p. 105874, Jun. 2025. doi: 10.1016/j.ijmedinf.2025.105874.

[7] Alamsyah and T. Fadila, "Increased accuracy of prediction hepatitis disease using PCA on a support vector machine," *J. Phys.: Conf. Ser.*, vol. 1968, no. 1, Jul. 2021. doi: 10.1088/1742-6596/1968/1/012016.

[8] L. Mochurad, V. Babii, Y. Boliubash, and Y. Mochurad, "Improving stroke risk prediction by integrating XGBoost, optimized PCA, and XAI," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, 2025. doi: 10.1186/s12911-025-02894-z.

[9] Y. Peng, S. Zhong, X. Li, F. Yu, and Z. Tang, "An artificial intelligence model to predict mortality among hemodialysis patients," *Comput. Biol. Med.*, vol. 168, p. 107764, Jan. 2025. doi: 10.1016/j.compbiomed.2025.107764.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[10] T. Zhou, Z. Ren, Y. Ma, L. He, J. Liu, J. Tang, and H. Zhang, "Early identification of bloodstream infection in hemodialysis patients by machine learning," *Heliyon*, vol. 9, no. 7, p. e18263, Jul. 2023. doi: 10.1016/j.heliyon.2023.e18263.

[11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002. doi: 10.1613/jair.953.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[15] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement," *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55–63, Jan. 2015. doi: 10.7326/M14-0697.

[16] G. S. Collins, K. G. M. Moons, J. A. Dhiman, S. G. F. S. R. Reitsma, R. D. Riley, and D. G. Altman, "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, Jan. 2019. doi: 10.7326/M18-1376.

[17] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[18] M. A. Loc, F. Renard, L. Fiedler, B. Marg, J. Renn, and D. Faller, "Catheter-Related Bloodstream Infections in Patients on Long-Term Hemodialysis: Epidemiology, Risk Factors, and Prevention Strategies," *J. Clin. Med.*, vol. 12, no. 5, p. 1793, Mar. 2023. doi: 10.3390/jcm12051793.

[19] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. A. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010. doi: 10.1097/EDE.0b013e3181c30fb2.