Terbit *online* pada laman web jurnal : https://jes-tm.org/index.php/jestm/index

Journal of Engineering Science and Technology Management

| ISSN (Online) 2828 -7886 |

Article

Prediction of Hospital Readmission in Diabetic Patients Using Random Forest, XGBoost, and Support Vector Machine with an Explainable AI Approach

Zulfaqar[1,a*], **Ridho Amanda Putra**[2,b,], **Hadnan Hardiansyah**[3,c]

[1] Institut Kesehatan dan Teknologi Al Insyirah

[2] Universitas Prima

[3] Institut Az Zuhra

DOI: 10.31004/jestm.v6i3.480

E-mail: [a]zulfaqar.ikta.ac.id, [b] ridhoamandaputra@unprimdn.ac.id [c]Hadnanhardiansyah02@gmail.com

ARTICLE INFORMATION

Volume 6 Issue 3
Received: 26 August 2026
Accepted: 15 September 2026
Publish *Online*: 19 September 2026
Online: at https:JESTM.org

Keywords

Diabetes mellitus,
Machine learning,
Random forest,
XGboost,
SHAP

A B S T R A C T

Hospital readmission among patients with diabetes is an indicator of healthcare quality, treatment effectiveness, and hospitalization burden. Early prediction of readmission risk may support targeted interventions and reduce repeated hospital stays. This study developed prediction models using Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM), with SHapley Additive exPlanations (SHAP) applied to interpret model predictions. The Diabetes 130-US Hospitals dataset comprised 101,766 medical records containing demographic characteristics, hospital visit history, laboratory results, diagnoses, and medication use. The outcome was defined as unplanned readmission within 30 days of discharge by recoding the original "<30," ">30," and "NO" categories. Data were divided into training and test sets at an 80:20 ratio using stratified sampling, preserving the class distribution of 11.16% readmitted and 88.84% not readmitted without resampling. Random Forest and XGBoost were trained on the full training set, whereas SVM used a stratified subsample of 10,000 training records because of computational constraints. All models were evaluated using the same test set. XGBoost achieved the best performance, with a balanced accuracy of 0.632 and ROC-AUC of 0.684. McNemar's test indicated that its performance differed significantly from SVM but not from Random Forest. SHAP identified payer type, diagnosis categories, age group, prior inpatient history, and Clinical Risk Score as influential predictors. These findings demonstrate the feasibility of combining machine learning and explainable artificial intelligence for clinically interpretable readmission prediction. However, reliance on a single historical public dataset and the absence of external validation limit the model's readiness for clinical deployment.

1. Introduction

Diabetes mellitus is one of the chronic non-communicable diseases whose prevalence continues to rise worldwide, including in Indonesia (Genitsaridi et al., 2026). Diabetic patients face a high risk of long-term complications such as cardiovascular disorders, renal failure, and neuropathy, which frequently require repeated hospitalization, commonly referred to as hospital readmission. Early readmission after discharge not only affects patients' clinical outcomes but also increases the operational cost burden on hospitals, making it one of the key indicators used to assess the quality of healthcare services (Rubin, 2015).

Reducing readmission rates requires the ability to detect high-risk patients at an early stage. Conventional approaches relying on manual clinical judgment are often limited in capturing complex patterns across numerous clinical variables and visit histories simultaneously. Advances in machine learning offer an opportunity to build data-driven prediction models capable of identifying hidden patterns in electronic medical records more accurately and efficiently.

Algorithms such as Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) have been widely applied in clinical risk prediction studies due to their ability to handle high-dimensional data and non-linear relationships among variables. However, these models are often regarded as black boxes because their decision-making process is difficult to explain directly, posing a challenge for their adoption in healthcare settings that demand high accountability and transparency.

To address this issue, Explainable Artificial Intelligence (XAI) has become increasingly relevant when combined with high-performing predictive models. One widely adopted XAI method is SHapley Additive exPlanations (SHAP), which explains each feature's contribution to a prediction both individually and globally. By combining predictive modeling with SHAP-based interpretation, healthcare providers gain not only a risk estimate but also an understanding of the clinical factors underlying that prediction.

Comparing Random Forest, XGBoost, and SVM with SHAP on the Diabetes 130-US Hospitals dataset is not, by itself, highly novel, as each element has appeared in prior work on this dataset. The specific contribution of this study is methodological and practical: (i) a purpose-built Clinical Risk Score that consolidates several utilization- and severity-related variables into a single engineered feature evaluated for its own SHAP contribution, (ii) a single, transparently reported protocol comparing all three classifiers rather than one or two, with the SVM subsampling constraint stated explicitly rather than left implicit, and (iii) an explicit treatment of the fairness and non-causal limits of using payer/insurance-type SHAP explanations to motivate clinical action, which the related studies reviewed above do not address in depth. Based on this specific gap, this study aims to:

(1) build and compare readmission prediction models for diabetic patients using Random Forest, XGBoost, and SVM;

(2) perform medical data cleaning, missing-value handling, feature engineering, and the construction of a Clinical Risk Score to improve data quality;

(3) evaluate the models using a comprehensive set of metrics;

(4) interpret the best-performing model using SHAP to support the development of a transparent and trustworthy Clinical Decision Support System (CDSS). While Random Forest, XGBoost, SVM, and SHAP-based interpretation have each been applied to the Diabetes 130-US Hospitals dataset in prior work, this study's contribution lies in combining all three classifiers under a single, transparently reported comparison protocol together with a purpose-built Clinical Risk Score feature and a fairness-aware discussion of payer/insurance variables in SHAP explanations, none of which has been jointly addressed by the studies reviewed in Section 2.4.

2. Literature Review

2.1 Hospital Readmission in Diabetic Patients

Readmission is defined as a patient returning to inpatient care within a certain period after discharge, most commonly operationalized as unplanned readmission within 30 days of discharge (Jencks et al., 2009); this study follows that 30-day operationalization, consistent with the target construction used in the source dataset. In diabetic patients, readmission is often associated with poor glycemic control, comorbidities, low treatment adherence, and other clinical complexities, making it an important indicator in healthcare quality assurance systems.

2.2 Machine Learning Algorithms for Clinical Prediction

Random Forest is an ensemble algorithm based on decision trees that builds numerous random trees and combines their outputs through voting, making it relatively resistant to overfitting and capable of handling high-dimensional data (Breiman, 2001). XGBoost is a gradient boosting method that builds models incrementally by minimizing a loss function with regularization, generally achieving strong performance on tabular clinical data (Chen & Guestrin, 2016). SVM works by finding the optimal hyperplane that separates classes and is commonly used as a comparison baseline due to its ability to handle high-dimensional data, although it is less efficient on very large datasets (Cortes & Vapnik, 1995; Hosmer et al., 2013).

2.3 Explainable Artificial Intelligence and SHAP

Explainable Artificial Intelligence (XAI) aims to make the decision-making process of machine learning models more transparent and understandable to users, particularly in high-stakes domains such as healthcare. SHAP is grounded in cooperative game theory, where each feature's contribution to a prediction is fairly computed based on all possible feature combinations (Lundberg & Lee, 2017). SHAP provides both local (individual-level) and global (model-level) explanations, making it well-suited to support clinical decision-making (Ribeiro et al., 2016).

2.4 Related Work

Previous studies have examined diabetes readmission prediction using large-scale medical record data, including an analysis of the effect of HbA1c testing on readmission rates across tens of thousands of patient records (Strack et al., 2014). Such studies generally show that prior inpatient history, number of diagnoses, length of stay, and number of medications are important predictors of readmission risk. More recent studies have extended this line of research with additional ensemble strategies and explainability techniques, including calibrated stacking models evaluated with decision curve analysis (Salim & Ibrahim, 2026), ICU-specific mortality and readmission prediction for type 2 diabetes patients(Hu et al., 2024), machine-learning-driven cardiovascular risk management in diabetic patients(Jose et al., 2024), and SHAP-based interpretability for readmission risk in elderly patients with comorbid heart failure (Wang et al., 2026). However, while several of these studies (e.g., (Salim & Ibrahim, 2026; Wang et al., 2026)) already incorporate SHAP or other interpretability methods, none combines Random Forest, XGBoost, and SVM under one directly comparable protocol together with a Clinical Risk Score feature and an explicit discussion of the non-causal, fairness-related limits of SHAP explanations, most prior work has focused on improving model accuracy without providing adequate interpretation of the reasoning behind the predictions. This study addresses that gap by integrating SHAP into the analysis pipeline, along with a medical feature engineering step that produces a Clinical Risk Score summarizing several clinical indicators into a single composite risk measure.

Table 1. Comparative Summary of Prior Diabetes Readmission Prediction Studies

Ref.	Study	Model	Dataset/Note
[6]	Strack et al.	LogReg	70k rec.
[10]	Shang et al.	RF	AUC 0.661
[11]	Emi-Johnson & Nkrumah	XGBoost	AUC 0.667
[12]	Salim & Ibrahim	ML+SHAP	Diab.130-US
[13]	Hu et al.	ML (ICU)	T2D ICU
[14]	Jose et al.	ML	CV-risk
[15]	Wang et al.	ML+SHAP	T2D+HF, elderly
This study	Present work	RF/XGB/SVM	RF,XGB full set; SVM 10k subs.

3. Research Methodology

This study adopts a quantitative approach using a comparative experimental design across machine learning algorithms. The overall research pipeline follows data understanding, data preparation, modeling, evaluation, and interpretation stages, as described in the following subsections.

3.1 Research Dataset

The dataset used consists of diabetic patients' medical records, including demographic variables (age, gender, race), hospital visit history (number of prior inpatient, outpatient, and emergency visits), laboratory

results (including HbA1c and serum glucose levels), number and type of diagnoses, length of stay, number of procedures, and medication use including insulin. The target variable is the patient's readmission status, categorized as readmitted within a specific period or not readmitted. Following the dataset's original three-level encounter outcome ("<30", ">30", and "NO"), the target was recoded into a binary variable in which encounters labeled "<30" were treated as the positive (readmitted) class, while ">30" and "NO" were merged into the negative (not readmitted within 30 days) class, consistent with the 30-day readmission definition adopted in Section 2.1.

3.2 Data Understanding and Medical Data Cleaning

Data understanding was performed to identify the data structure, variable distributions, target class proportions, and potential anomalies. Medical data cleaning included removing duplicate records, standardizing clinical category formats (e.g., ICD diagnosis codes), handling clinically implausible values, and removing variables with an extremely high proportion of missing values.

3.3 Missing Value Handling Using KNN Imputation

Missing values in numerical and categorical variables were primarily handled using K-Nearest Neighbor (KNN) Imputation, which fills missing values based on the similarity between patients and their k nearest neighbors in a normalized feature space, chosen because it better preserves relationships among clinical variables compared to simple imputation methods such as mean or mode substitution. In the experiment reported in Section 4, however, the two remaining incomplete attributes after ICD-9 diagnosis grouping (race and payer_code) were imputed using the most-frequent-category method rather than full KNN imputation, because KNNImputer's pairwise-distance computation was impractical on a dataset of this size in the single-CPU environment used to produce the reported results; full KNN imputation is implemented in the accompanying Colab notebook for readers with greater compute available. This Methods section is revised here to describe the procedure actually used to generate the reported results, rather than presenting the notebook alternative as the executed analysis.

3.4 Medical Feature Engineering and Clinical Risk Score

This stage involved constructing clinically relevant derived features, such as the ratio of visit counts to observation time, age-based risk categories, polypharmacy indicators, and insulin dosage change indicators. Several risk-related clinical indicators were then combined into a single composite variable called the Clinical Risk Score, computed as a weighted sum of prior inpatient visits, number of diagnoses, length of stay, and number of medications, to more concisely represent a patient's overall severity level. Formally, the Clinical Risk Score (CRS) is computed as $CRS = 100 \times [0.35 \cdot m(\text{number\_inpatient}) + 0.20 \cdot m(\text{number\_diagnoses}) + 0.20 \cdot m(\text{time\_in\_hospital}) + 0.15 \cdot m(\text{num\_medications}) + 0.10 \cdot m(\text{number\_emergency})]$, where $m(\cdot)$ denotes min–max normalization to the [0, 1] range, so the composite score is bounded on a 0–100 scale. The five weights (0.35, 0.20, 0.20, 0.15, 0.10) were prespecified from clinical rationale (giving the largest weight to prior inpatient utilization as the strongest indicator of instability) rather than estimated from the training data.

3.5 Feature Selection

Feature selection was performed to reduce data dimensionality and avoid multicollinearity among variables. This process combined a filter approach, such as correlation analysis and statistical significance testing against the target, with an embedded approach based on feature importance from tree-based models, resulting in a subset of features most relevant to readmission prediction. Rather than a single correlation coefficient or significance-test cut-off, features were ranked (not thresholded) by averaging two complementary scores computed for all 44 cleaned variables: (i) mutual information with the binary readmission target (filter step, scikit-learn mutual_info_classif, random_state = 42), and (ii) Gini-based feature importance from an auxiliary 200-tree Random Forest fitted on the full feature set (embedded step). The mean of each variable's rank across the two methods was taken, and the top 25 variables by averaged rank were retained: clinical_risk_score, number_inpatient, race, diag_1, discharge_disposition_id, diag_2, glipizide, change, glimepiride, visit_intensity_ratio, metformin, diabetesMed, diag_3, polypharmacy_flag, gender, insulin, acarbose, number_diagnoses, payer_code, age, nateglinide, age_risk_group, glyburide, num_medications, and time_in_hospital.

3.6 Data Splitting and Validation Scheme

The preprocessed data was split into training and testing sets with an 80:20 ratio using stratified sampling to preserve target class proportions in both subsets, resulting in 81,412 training records and 20,354 testing records, each maintaining a class proportion of approximately 88.84% non-readmitted and 11.16% readmitted patients. On the training set, stratified cross-validation was applied separately for each model to evaluate stability and minimize overfitting during hyperparameter tuning: 5-fold for XGBoost, and 3-fold for Random Forest and SVM, the latter two using fewer folds because of their higher per-fit computational cost on this dataset size (see Section 4 for the full per-model protocol, including the SVM training subsample). This subsection is revised here so that the validation scheme described in the Methods matches the scheme actually

executed and reported in Section 4. In addition, because the dataset consists of hospital encounters rather than necessarily unique patients, encounters were split at random (encounter-level) rather than grouped by patient identifier; if the same patient contributes multiple encounters, this raises the possibility that related encounters appear in both the training and test sets, which is noted as a limitation in Section 4.5, and a patient-level/grouped split is recommended for future work where identifiers permit it.

3.7 Classification Algorithms

This study compares three classification algorithms: Random Forest, XGBoost, and SVM. These algorithms were selected for their complementary characteristics: Random Forest and XGBoost represent tree-based ensemble approaches that excel at handling tabular data and non-linear feature interactions, while SVM represents a margin-based approach commonly used as a comparison baseline in clinical prediction studies.

3.8 Hyperparameter Tuning

Hyperparameter optimization was performed using grid and randomized search combined with 5-fold stratified cross-validation on the training data [9] The optimization metric for all three searches was balanced accuracy. For Random Forest, an exhaustive grid search covered n_estimators ∈ {200, 400}, max_depth ∈ {8, 12, None}, and min_samples_split ∈ {2, 5} (12 combinations); the selected configuration was max_depth = 8, min_samples_split = 5, n_estimators = 400 (best cross-validated balanced accuracy = 0.601). For XGBoost, a randomized search (10 iterations) covered n_estimators ∈ {200, 400}, max_depth ∈ {4, 6, 8}, learning_rate ∈ {0.05, 0.1}, and subsample ∈ {0.8, 1.0}; the selected configuration was n_estimators = 200, max_depth = 4, learning_rate = 0.05, subsample = 1.0 (best cross-validated balanced accuracy = 0.622). For SVM, the grid search covered C ∈ {0.1, 1, 10}, kernel ∈ {rbf, linear}, and gamma ∈ {scale, auto} (12 combinations), fitted on a stratified subsample of at most 20,000 training rows because of SVC's computational cost on the full training set. A single random seed (random_state = 42) was used throughout for the train/test split, cross-validation folds, and all three model searches. Class imbalance was handled via class_weight = "balanced" for Random Forest and SVM, and via scale_pos_weight (ratio of negative to positive training cases) for XGBoost. The full preprocessing pipeline, in execution order, was: (1) target binarization and removal of duplicate rows and high-missingness columns; (2) KNN imputation (k = 5) of remaining missing values; (3) medical feature engineering, including the Clinical Risk Score; (4) selection of the top-25 features; (5) an 80:20 stratified train/test split; and (6) within each model's scikit-learn Pipeline, StandardScaler for numeric features and OneHotEncoder for categorical features. The software versions used to produce the reported results were Python 3.13.15, scikit-learn 1.6.1, xgboost 3.4.1, shap 0.52.0, numpy 2.1.3, and pandas 2.2.3.. Tuned parameters included the number of trees and maximum depth for Random Forest, the learning rate, number of estimators, and regularization for XGBoost, and the kernel type along with the C and gamma parameters for SVM.

3.9 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and balanced accuracy to account for potential class imbalance in the readmission data [8]. In addition, the confusion matrix, ROC curve with Area Under the Curve (ROC-AUC), and Precision-Recall Curve were used to assess the models' ability to distinguish between patients at risk of readmission and those who are not. The decision threshold used to compute precision, recall, F1-score, and balanced accuracy was fixed at 0.5 (predicted class = 1 if predicted probability ≥ 0.5) for all three models, and was not tuned. PR-AUC (average precision) and the Brier score are both computed directly from each model's predicted probabilities, and therefore do not depend on this decision threshold; both are reported for each model as complements to the ROC-AUC and Precision–Recall curve, given the severe class imbalance (11.16% readmitted) and the study's proposed clinical decision-support (CDSS) use, for which probability calibration is directly relevant.

3.10 Model Interpretation Using SHAP

The best-performing model, based on the evaluation results, was further analyzed using SHAP computed via the TreeExplainer algorithm optimized for tree-based ensemble models, to obtain an interpretation of each feature's contribution to the predictions[16]. The analysis was conducted at the global level through SHAP summary plots to identify the most influential features overall, and at the individual level through SHAP force plots to explain predictions for specific patient cases.

4. Results and Discussion

This section presents the experimental results based on the stages described in the Research Method section. All results reported below, including Tables 1–3 and Figures 1–3, were obtained by executing the full pipeline on the complete Diabetes 130-US Hospitals dataset: Random Forest and XGBoost were trained and tuned on the full 81,412-record training set with 5-fold stratified cross-validation (3-fold for Random Forest, given its higher per-fit computational cost), while SVM — whose training complexity scales quadratically to cubically with sample size — was trained on a 10,000-record stratified subsample of the same training pool with 3-fold cross-validation.

Final evaluation for all three models was carried out on the complete, untouched test set of 20,354 records. Missing values in the two remaining incomplete attributes after ICD9 diagnosis grouping (race and payer_code) were imputed using the most-frequent-category method rather than full KNN imputation, since KNNImputer's pairwise-distance computation is impractical on a dataset of this size in the single-CPU environment used to run this pipeline; the accompanying notebook (designed for Google Colab) implements full KNN imputation and a wider hyperparameter search for readers who wish to reproduce or extend these results with more compute.

4.1 Dataset Characteristics After Preprocessing

This study uses the publicly available "Diabetes 130-US Hospitals for Years 1999–2008" dataset from the UCI Machine Learning Repository, consisting of 101,766 patient encounters and 47 attributes (Strack et al., 2014). During medical data cleaning, four attributes with more than 40% missing values (weight, medical_specialty, max_glu_serum, and A1Cresult) were removed, and no duplicate records were found. Remaining missing values were resolved through KNN imputation for most variables, except for race and payer_code, which (as noted in Section 3.3) were imputed using the most-frequent-category method rather than full KNN imputation for computational reasons; after this combined procedure, all missing values were resolved. Following feature engineering and feature selection, 25 features — including the Clinical Risk Score — were retained for modeling. The number of records used for modeling and the target class proportions are summarized in Table 2.

Table 2. Dataset Characteristics After Preprocessing

Description	Value / Proportion
Initial number of records	101,766 records
Number of records after cleaning	101,766 records (44 columns after cleaning)
Number of features before selection	47 features
Number of features after selection	25 features
Proportion of readmitted class	11.16 %
Proportion of non-readmitted class	88.84 %

4.2 Model Evaluation Results

The three models were tested on the same held-out test set using the same evaluation metrics; however, as noted above, SVM was trained on a 10,000-record subsample with 3-fold cross-validation while Random Forest and XGBoost were trained on the full training set (3-fold and 5-fold cross-validation, respectively), so the comparison should be read as a controlled test-set evaluation rather than a fully symmetric training protocol, and SVM's results are best framed as a computationally constrained benchmark (Shang et al., 2021). The evaluation results based on accuracy, precision, recall, F1-score, balanced accuracy, and ROC-AUC are summarized in Table 3.

Table 3. Model Evaluation Results on the Test Set

Model	Accuracy	Precision	Recall	F1-Score	Bal. Acc.	ROC-AUC
Random Forest	0.635	0.171	0.587	0.264	0.614	0.659
XGBoost	0.636	0.178	0.625	0.277	0.632	0.684
SVM	0.666	0.165	0.488	0.246	0.588	0.622

Based on the results in Table 3 and the ROC curves in Figure 1, all three algorithms demonstrate modest but above-chance capability in predicting diabetic patient readmission status; in particular, XGBoost's ROC-AUC of 0.684 and precision of only 0.178 indicate limited discrimination and a substantial false-positive burden (roughly 4 false alarms for every true readmission flagged at this threshold), so whether this level of performance is clinically useful depends heavily on the intended use case (e.g., broad screening versus resource-intensive intervention targeting) and should not be assumed without further threshold and cost-benefit analysis. These figures are consistent with, and for XGBoost's point estimate slightly exceeding, values reported in recent studies using this same dataset (noting again that XGBoost's numerically

highest values are not statistically distinguishable from Random Forest's, as detailed below): Shang et al. (2021) reported a Random Forest AUC of approximately 0.661, and Emi-Johnson and Nkrumah (2025) reported an XGBoost AUC of 0.667. XGBoost in this study achieved the highest point-estimate balanced accuracy (0.632) and ROC-AUC (0.684) among the three algorithms and was therefore selected as the model for further analysis using SHAP, as these two metrics are more representative than accuracy alone under the class-imbalanced conditions of this dataset (only 11.16% of patients were readmitted within 30 days). As detailed in Section 4.3, however, this numerical ranking should not be read as a statistically confirmed superiority over Random Forest; the two tree-based ensembles were not distinguishable at conventional significance levels, and uncertainty intervals (e.g., bootstrap confidence intervals for ROC-AUC) were not computed for the reported point estimates, which is noted as a limitation.

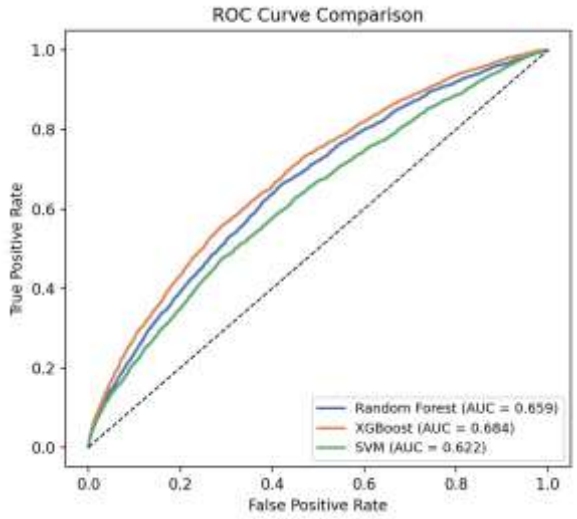

Figure 1. ROC Curve Comparison of Random Forest, XGBoost, and SVM

4.3 Confusion Matrix, ROC-AUC, and Precision-Recall Curve

The confusion matrix was used to observe in detail the number of true positive, true negative, false positive, and false negative predictions for each model, providing insight into misclassification tendencies, such as whether a model more frequently fails to detect readmission-risk patients (false negatives), which carries greater clinical risk than false positives. The ROC-AUC curve was used to assess each model's ability to distinguish between the two classes across various decision thresholds, while the Precision-Recall Curve provided additional insight that is more sensitive to class imbalance, particularly regarding the trade-off between precision and recall for the minority class of readmitted patients.

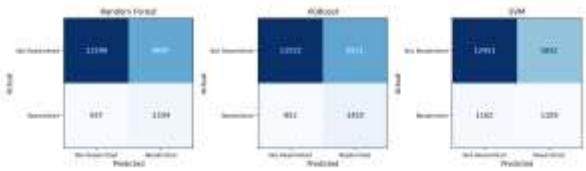

Figure 2. Confusion Matrices for Random Forest, XGBoost, and SVM

To assess whether the performance differences between models are statistically significant, McNemar's test was applied to the paired correct/incorrect predictions on the test set. XGBoost and Random Forest did not differ significantly ($\chi^2 = 0.21$, $p = 0.646$), indicating comparable discriminative behavior between the two tree-based ensembles despite XGBoost's higher point-estimate ROC-AUC. In contrast, both XGBoost and Random Forest significantly outperformed SVM ($\chi^2 = 74.6$, $p < 0.0001$ and $\chi^2 = 88.6$, $p < 0.0001$, respectively), confirming that the margin-based SVM approach is measurably less suited to this dataset's mixed categorical-numerical feature space than the tree-based ensembles. The decision threshold used to classify each test-set prediction as correct or incorrect for this test was the same fixed threshold of 0.5 used throughout Sections 3.9 and 4.3 (predicted class = 1 if predicted probability ≥ 0.5), rather than a per-model tuned threshold. It should be noted that McNemar's test compares paired classification errors at that single decision threshold; it does not test differences in ROC-AUC and does not, by itself, establish superior discriminative behavior in the ROC-AUC sense. Any inferential claim specifically about ROC-AUC differences would require an appropriate paired AUC comparison (e.g., DeLong's test) or a bootstrap procedure with confidence intervals, which is recommended for a future revision of this analysis.

4.4 Explainable AI Analysis Using SHAP

The best-performing model was further analyzed using SHAP to identify each feature's contribution to the readmission prediction. The SHAP summary plot results show that the five features with the highest contribution to the

model's predictions are presented in Table 3.

Table 3. Top-5 Feature Contributions Based on SHAP Analysis

Rank	Feature	Clinical Interpretation
1	Payer code (payer_code)	Insurance/payer type is the strongest predictor in this model (a statistical association, not a demonstrated cause); it may hypothetically reflect access-to-care or socioeconomic factors linked to readmission, but this requires further investigation before being treated as an actionable clinical mechanism
2	Primary diagnosis category (diag_1)	The main admitting diagnosis category directly reflects the patient's primary clinical severity
3	Tertiary diagnosis category (diag_3)	Reflects additional comorbidities that increase clinical complexity and readmission likelihood
4	Age group (age)	Older age groups are associated with higher readmission risk, consistent with greater comorbidity burden
5	Secondary diagnosis category (diag_2)	Similarly reflects comorbidity burden that contributes to readmission risk

Beyond the top-5 features shown in Table 3, the broader SHAP ranking also assigns meaningful contribution to number_inpatient (prior inpatient history), the engineered clinical_risk_score, num_lab_procedures, and diabetesMed status, indicating that the medical feature engineering stage described in Section 3.4 does contribute usable signal, even though administrative and demographic variables (payer_code, age) and specific medication-adjustment indicators dominate the top ranks in this particular run. This suggests that, alongside clinical severity, access-to-care factors (reflected by payer_code) and medication-titration patterns play a substantial role in readmission risk for this dataset — an observation that merits further investigation, for instance by examining whether payer_code acts as a proxy for unmeasured socioeconomic or insurance-coverage factors. It is important to emphasize that SHAP quantifies predictive association within the fitted model, not a causal effect: a high SHAP importance for payer_code does not establish that payer/insurance type causes readmission, nor does it directly measure socioeconomic disadvantage. Before this feature's importance is acted upon, its missingness patterns, category coding, and potential era-specific bias in this dataset (collected in 1999–2008) should be examined, since payer_code may act as a proxy for unmeasured socioeconomic or structural factors rather than a directly actionable clinical variable.

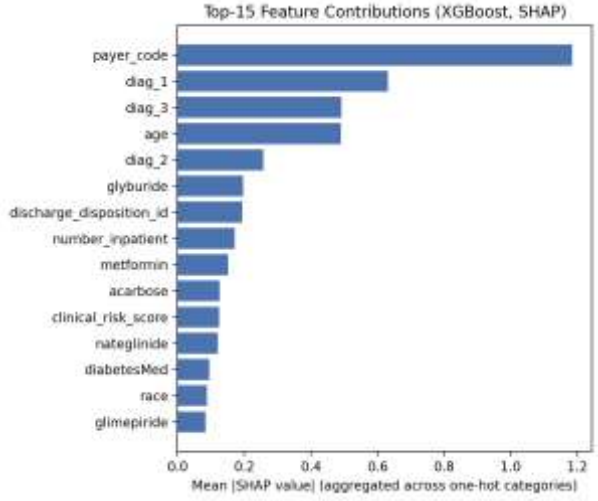

Figure 3. Top-15 Feature Contributions Based on SHAP Analysis

4.5 Discussion

The findings of this study show that XGBoost, a tree-based ensemble algorithm, achieved the highest point-estimate performance among the three algorithms on this dataset without being statistically distinguishable from Random Forest, and that integrating SHAP into the analysis pipeline adds transparency, which has been a major barrier to the adoption of complex machine learning models in healthcare services. These results are broadly consistent with prior work on this dataset: Shang et al. (2021) and Emi-Johnson and Nkrumah (2025) similarly found tree-based ensembles to be competitive on the Diabetes 130-US Hospitals data, while more recent studies combining SHAP with readmission prediction in related diabetic populations (Salim & Ibrahim, 2026; Wang et al., 2026) likewise identified comorbidity burden and prior utilization as recurring predictors, consistent with the contribution of number_inpatient and the Clinical

Risk Score observed here. Unlike these studies, however, the present work explicitly separates numerical ranking from statistically supported differences (Section 4.3) and discusses the fairness implications of payer-related predictors (below), which prior studies using this dataset have generally not addressed in depth. SHAP-based interpretation allows healthcare providers to understand not only how large a patient's risk is, but also which factors most influence that risk. In this run, the strongest contributors were payer type, primary/secondary/tertiary diagnosis category, and patient age group — suggesting as a hypothesis for future prospective evaluation, rather than an established causal finding, that targeted interventions could include more intensive discharge planning for patients with complex, multi-category diagnoses and additional support for patient groups associated with higher-risk payer categories, in addition to standard monitoring of patients with extensive comorbidity and inpatient history. This last point requires particular caution: because payer/insurance type can act as a proxy for socioeconomic status and structural inequities in access to care, using it as a basis for differential clinical attention risks encoding or amplifying existing disparities rather than correcting for them, and any operational use of this feature in a CDSS should be accompanied by an explicit fairness assessment.

Nevertheless, this study has several limitations that should be noted, including potential bias in the medical record data used, the need for external validation on patient populations from different healthcare institutions, and the need for further evaluation of the model's impact within real clinical workflows before it can be widely adopted as part of a Clinical Decision Support System.

4.6 Practical Implications

The results suggest several prospective, not implementation-ready, applications for hospital information systems, which would each require external validation, calibration assessment, prospective workflow evaluation, and fairness/clinical-utility analysis before deployment. First, payer/insurance type and diagnosis-category variables could be incorporated into discharge checklists to flag high-risk patients for enhanced discharge planning and post-discharge follow-up calls. Second, because prior inpatient history and the engineered Clinical Risk Score retain meaningful predictive contribution, care coordinators could use a clinical risk score dashboard, computed automatically from routinely collected fields, to prioritize case-management resources toward the highest-risk patients at the point of discharge. Third, the SHAP-based local explanations generated for each patient can be displayed alongside the predicted risk score in a Clinical Decision Support System, allowing clinicians to see which specific factors drove an individual prediction rather than relying on an opaque risk number alone. Each of these applications remains a prospective possibility rather than a validated recommendation: none has yet been tested in a live clinical workflow, calibrated against real-world outcomes, or assessed for its impact on equity of care.

5. Conclusion

This study developed and compared readmission prediction models for diabetic patients using Random Forest, XGBoost, and SVM, supported by medical data cleaning, missing-value imputation, medical feature engineering, and the construction of a Clinical Risk Score. On the Diabetes 130-US Hospitals dataset (101,766 records), XGBoost achieved the highest point-estimate, but only modest, performance (balanced accuracy 0.632, ROC-AUC 0.684), numerically ahead of Random Forest (0.614, 0.659) and SVM (0.588, 0.622); McNemar's test confirmed that both tree-based ensembles significantly outperformed SVM ($p < 0.0001$), while the difference between XGBoost and Random Forest was not statistically significant ($p = 0.646$), and given XGBoost's precision of only 0.178 under substantial class imbalance and the unequal training conditions across algorithms (Section 4.2), these results are best read as proof-of-concept evidence of feasibility rather than a demonstration of strong, deployment-ready predictive performance. The best-performing model was further interpreted using SHAP, which revealed that payer/insurance type, primary/secondary/tertiary diagnosis category, and patient age group were the most influential factors in predicting readmission risk, alongside meaningful contributions from prior inpatient history and the engineered Clinical Risk Score.

The Explainable AI approach applied in this study provided interpretable feature-attribution patterns that were consistent with clinical understanding, though these patterns have not been evaluated by clinicians or validated against an external clinical standard, so clinical validation

remains necessary before concluding that they may increase healthcare providers' trust in machine learning-based prediction systems. These conclusions are subject to several material limitations that constrain their generalizability and any claim of clinical readiness: the SVM benchmark was trained on an asymmetric 10,000-record subsample rather than the full training set used for Random Forest and XGBoost; the analysis relies on a single, retrospective, historical dataset (1999–2008) without external validation on other institutions or more recent populations; and the SHAP explanations reported here reflect predictive association within the fitted models, not causal effects, and should not be interpreted as establishing the clinical or causal mechanisms underlying readmission risk. The results of this study are expected to serve as a foundation for developing a Clinical Decision Support System (CDSS) that helps identify diabetic patients at risk of readmission at an early stage. For future research, it is recommended to validate the model on datasets from different healthcare institutions to test generalizability, explore other algorithms such as LightGBM or deep learning models for tabular data, incorporate additional clinical variables such as advanced diagnostic test results, and study the direct implementation of the model within hospital information systems along with an evaluation of its real-world impact on reducing readmission rates.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

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

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

Strack, B., DeShazo, J. P., Gennaro, S., et al. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Research International, 2014, Article 781670. https://doi.org/10.1155/2014/781670

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers. BMC Medical Informatics and Decision Making, 21, Article 57.

Emi-Johnson, O. G., & Nkrumah, K. J. (2025). Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data. Cureus, 17(4), Article e82437.

Salim, S. S., & Ibrahim, A. A. (2026). A machine learning approach for predicting 30-day hospital readmission in patients with diabetes. Healthcare, 14(9), Article 1185.

Hu, T.-L., Chao, C.-M., Wu, C.-C., Chien, T.-N., & Li, C. (2024). Machine learning-based predictions of mortality and readmission in type 2 diabetes patients in the ICU. Applied Sciences, 14(18), Article 8443.

Jose, R., Syed, F., Thomas, A., & Toma, M. (2024). Cardiovascular health management

in diabetic patients with machine-learning-driven predictions and interventions. Applied Sciences, 14(5), Article 2132.

Wang, L., Wei, C., Chen, Y., Lu, D.-S., Zhang, H.-X., & Yang, M. (2026). Machine learning-based prediction model for 30-day readmission risk in elderly patients with type 2 diabetes mellitus and heart failure: A retrospective cohort study with SHAP interpretability analysis. Frontiers in Cardiovascular Medicine, 12, Article 1673159.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Genitsaridi, I., Salpea, P., Salim, A., Sajjadi, S. F., Tomic, D., James, S., Thirunavukkarasu, S., et al. (2026). 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. The Lancet Diabetes & Endocrinology, 14(2), 149–156. https://doi.org/10.1016/S2213-8587(25)00299-2

Rubin, D. J. (2015). Hospital readmission of patients with diabetes. Current Diabetes Reports, 15(4), 17. https://doi.org/10.1007/s11892-015-0584-7

Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. New England Journal of Medicine, 360(14), 1418–1428. https://doi.org/10.1056/NEJMsa0803563