

Explainable, Non-Invasive Prediction of Diabetes Hospital Readmission: A SHAP-Interpretable Stacking Ensemble for Edge-Deployable Industry 5.0 Healthcare

Aminu Usman Jibril¹, A. Senthil Kumar², Abdullahi Shehu Ahmad³

¹Department of Computer Science, Skyline University, Nigeria

²Dean, School of Science and Information Technology (SSIT), Associate Professor, Department of Computer Science, Skyline University Nigeria

³Computer Science Department, Faculty of Computing, Bayero University, Kano

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Diabetes mellitus remains a major global health challenge with hospital readmission that served as an important measure of healthcare quality and patient outcomes. This study presents an explainable hybrid artificial intelligence framework for non-invasive diabetes readmission prediction using multimodal patient data and edge-computing capabilities aligned with Industry 5.0 principles. A stacking ensemble model that combine XGBoost, LightGBM, ExtraTrees and logistic regression was developed using 40,000 hospital encounters from the Diabetes 130-US Hospitals dataset. The proposed framework achieved a ROC-AUC of 0.7042 and Average Precision of 0.6678 which outperforming several baseline models. To enhance transparency and clinical trust the SHAP-based explainability was integrated to provide both global and local interpretation of predictions. Feature engineering using non-invasive administrative data effectively identified readmission risk patterns with readmitted patients showing significantly higher prior visit counts and clinical risk scores. The framework was further designed with edge deployment in mind which favoured lightweight compression-compatible models and architectures with low predicted inference latency with the goal of enabling real-time privacy-preserving prediction in resource-constrained healthcare environments pending on-device validation. This work supports Industry 5.0 goals of human-centric resilient and sustainable healthcare that enable transparent AI-driven decision-making without reliance on invasive laboratory tests. The study also demonstrates the strong performance of XGBoost on structured clinical data and highlights the limited performance gains achieved through the stacking ensembles.

Keywords: Diabetes prediction; Explainable AI; Hybrid machine learning; Edge computing; Industry 5.0; Healthcare analytics; Hospital readmission; Stacking ensemble; SHAP; XGBoost

Online publication: October 23, 2026

1. Introduction

1.1 Context and motivation

The convergence of artificial intelligence (AI) and healthcare represents a transformative opportunity for chronic disease management. Diabetes mellitus affects over 537 million adults globally, with projections to reach 643 million by 2030 and 783 million by 2045 ^[1]. Hospital readmissions among diabetic patients impose substantial economic burdens which was estimated at USD 966 billion globally as of 2021 and this indicate critical gaps in discharge planning that follow-up care coordination and patient education ^[2].

Industry 5.0 is characterized by human-machine collaboration, resilience and sustainability that provides a framework for reimagining healthcare delivery. Unlike Industry 4.0's that focus on automation and efficiency. Industry 5.0 emphasises human-centricity and placing clinicians and patients at the center of AI-assisted decision-making ^[3]. This paradigm shift demands AI systems that are not only accurate but also interpretable, trustworthy, privacy-preserving and deployable in diverse clinical environments from well-equipped urban hospitals to resource-constrained rural clinics.

1.2. Problem statement

Despite robust technological advances, the early and reliable prediction of diabetes readmission remains beset by persistent challenges. Three critical gaps motivate this research:

Most state-of-the-art AI models for diabetes prediction operate as “black boxes,” delivering high accuracy but offering little insight into their decision-making processes ^[4]. In healthcare applications where decisions have life-altering consequences this lack of transparency limits clinician trust and impedes regulatory compliance (e.g., GDPR's “right to explanation”) and constrains ethical deployment.

Traditional diagnostic paradigms rely heavily on invasive laboratory biomarkers (fasting plasma glucose, HbA1c, oral glucose tolerance tests) that require specialized equipment, trained personnel and stable supply chains resources that are often unavailable in low- and middle-income countries (LMICs) where 75% of the global diabetes burden resides ^[1].

While cloud-based AI offers powerful computational capabilities, real-time readmission prediction requires low-latency inference at the point of care. Existing frameworks rarely address model compression, privacy preservation, and offline functionality essential for resource-constrained clinical settings ^[5].

1.3. Research objectives and questions

This research aims to design, develop, and validate an integrated explainable and edge-deployable AI framework for predicting diabetes hospital readmission using multimodal non-invasive patient data. Specific research questions include:

- (1) To what extent can a hybrid ML/DL architecture using non-invasive data achieve screening-grade prediction accuracy ($\text{ROC-AUC} \geq 0.70$, the conventional threshold for acceptable discrimination) compared to traditional biomarker-based approaches?
- (2) How can explainable AI mechanisms (SHAP, LIME) be integrated to provide transparent, clinically actionable predictions without compromising accuracy?
- (3) Can the model be optimized for edge deployment through algorithmic efficiency and model compression while maintaining predictive performance?

1.4. Contributions and paper organization

The key contributions of this paper are:

- (1) A hybrid stacking ensemble architecture achieving ROC-AUC of 0.704 on a benchmark diabetes readmission dataset;
- (2) Integration of SHAP for global and local interpretability;
- (3) A comprehensive feature engineering framework capturing clinical risk from non-invasive administrative data;
- (4) Analysis of precision-recall trade-offs for clinical deployment. The remaining part of this paper is organised as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 presents results; Section 5 discusses findings and Industry 5.0 alignment; Section 6 concludes and outlines future work.

2. Literature review

2.1. Evolution of AI in diabetes prediction

The application of AI to diabetes prediction has progressed through three distinct phases. Early work (1980s–2000s) employed statistical methods logistic regression and decision trees on small, structured datasets like the Pima Indian Diabetes Dataset (PIDD) that achieved modest predictive accuracy but offering interpretability^[6]. The second phase (2000–2015) saw adoption of machine learning algorithms including support vector machines (SVM), random forests (RF), and k-nearest neighbors (k-NN), that demonstrated an improved performance on heterogeneous clinical data^[7].

The current phase (2015–present) is characterized by deep learning architectures convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks and transformers that automatically learn hierarchical feature representations from raw data. Hybrid CNN-LSTM models have proven particularly effective for mining temporal patterns in electronic health records (EHRs) and biomedical signals^[8, 9]. However, these models’ “black- box” nature has motivated complementary research in explainable AI.

2.2. Explainable AI in healthcare

Explainable AI (XAI) has emerged as critical for bridging the gap between predictive performance and clinical trust. Post-hoc methods applied after model training dominate current practice. SHAP (Shapley Additive Explanations) provides theoretically sound feature attribution based on cooperative game theory that assign each feature an importance score for individual predictions^[10]. LIME (Local Interpretable Model-Agnostic Explanations) approximates complex model decisions locally with interpretable surrogate models^[11].

Recent hybrid approaches combine prediction and interpretability. Nguyen et al. demonstrated that HyperTab-LIME achieves meaningful gains in both accuracy and interpretability for prediabetes screening^[12]. Regulatory frameworks increasingly mandate explainability: the European Union’s GDPR includes provisions for a “right to explanation,” while the FDA has proposed specific requirements for AI/ML-based Software as a Medical Device (SaMD)^[13].

2.3. Edge computing for diabetes prediction

Edge computing processing data near its source rather than in centralised clouds offers compelling

advantages for healthcare: reduced latency (critical for real-time alerts), enhanced privacy (sensitive data stays local) and offline functionality (essential for rural clinics with unreliable connectivity)^[14]. Frameworks like HealthEdge and SmartEdge demonstrate the feasibility of ML-based diabetes prediction on IoT-edge-cloud infrastructure^[15,16].

However, edge devices have limited computational resources. Model compression techniques pruning (removing redundant weights), quantisation (reducing numerical precision), and knowledge distillation (training smaller “student” models) address these constraints. Han et al. showed that deep networks can be pruned by over 80% with minimal accuracy loss^[17]. Lightweight architectures (MobileNet, ShuffleNet, EfficientNet) are increasingly adopted for mobile health applications^[18,19].

2.4. Critical gaps identified

Our systematic literature synthesis reveals three underexplored areas:

- (1) Explainability in hybrid deep models remains underdeveloped with most studies relying on post-hoc methods rather than intrinsically interpretable designs;
- (2) Edge deployment of explainable hybrid models is rare, with existing frameworks focusing on traditional ML ensembles rather than deep architectures;
- (3) Few frameworks target readmission prediction (rather than disease onset), leaving a gap for solutions integrating temporal modelling, multimodal fusion, and real-time inference to improve hospital resource planning and patient outcomes^[20,21].

3. Methodology

3.1. Dataset description

This study utilized the Diabetes 130-US Hospitals dataset a large de-identified repository originally comprising 101,766 hospital encounters spanning ten years, collected from 130 hospitals and integrated delivery networks across the United States^[22]. A stratified random subsample of 40,000 encounters was drawn from the full repository to balance computational tractability during model development and hyperparameter search against statistical power; stratification preserved the original class proportions, and preliminary checks confirmed that distributions of the principal demographic and utilization variables in the subsample did not differ materially from the full dataset. This subsampling step is a design choice rather than a structural property of the data, and validation on the full repository remains a useful robustness check for future work. The dataset includes 26 attributes across seven categories: demographic (race, gender, age), administrative (admission type, discharge disposition, admission source), clinical utilization (time in hospital, number of laboratory procedures, number of procedures, number of medications), visit history (outpatient, emergency, and inpatient visits in the past year), diagnostic (primary, secondary, and tertiary ICD-9 codes), laboratory (maximum glucose serum level, A1C result) and medication indicators (metformin, insulin, medication change, diabetes medication). All records were de-identified prior to public release by the original data custodians in accordance with HIPAA Safe Harbor provisions as a secondary analysis of a publicly available, de-identified dataset no additional institutional ethics approval was required and no attempt was made to re-identify individual patients or institutions.

The target variable readmitted originally contained three categories: ‘No’ (53.18%), ‘> 30 days’ (35.59%), and ‘< 30 days’ (11.23%). For binary classification, we mapped ‘No’ to 0 (non-readmitted) and both ‘> 30

days' and '< 30 days' to 1 (readmitted). The final target distribution comprised 21,273 negative cases (53.18%) and 18,727 positive cases (46.82%), yielding a class imbalance ratio of approximately 1.14:1. This near-balanced distribution is considerably more favorable for classification than the severe imbalance typical of disease-onset prediction tasks, reducing though not eliminating the risk of majority-class bias without requiring resampling techniques such as SMOTE.

3.2. Data preprocessing

Variables with excessive missingness (> 80%) max_glu_serum (94.72%) and A1Cresult (83.11%) were excluded to avoid substantial imputation bias. For categorical variables with manageable missingness (diagnostic categories), mode imputation was employed using the most frequent category. For numerical variables, median imputation was applied, which is robust against outliers and preserves central tendency.

ICD-9 diagnosis codes (diag_1, diag_2, diag_3) were systematically mapped to 16 clinically meaningful disease categories following standard ICD-9 classification schemes infectious, neoplasms, endocrine, blood, mental, nervous, circulatory, respiratory, digestive, genitourinary, skin, musculoskeletal, injury, supplemental, external, and other. This transformation reduced dimensionality while preserving diagnostic information.

Categorical variables were label-encoded (rather than one-hot encoded) because tree-based models (XGBoost, LightGBM, Random Forest) effectively handle ordinal encoded categorical variables without excessive dimensionality expansion^[23]. Numerical features were standardized using z-score normalization to ensure comparable scales: $z = (x - \mu) / \sigma$, where μ is the feature mean and σ is the feature standard deviation. The preprocessed feature set comprised 44 numerical features with no remaining missing values.

3.3. Feature engineering

Twenty derived clinical risk indicators were engineered to capture patient acuity healthcare utilization intensity, medication management complexity and demographic risk factors. Key engineered features are summarized in **Table 1**.

The clinical risk score weights were designed based on clinical domain knowledge. Inpatient visits (0.4) and emergency visits (0.3) as the strongest predictors which are followed by high medication burden (0.2) and high diagnostic burden (0.1). This weighting was informed by prior research demonstrating that prior healthcare utilization is among the strongest predictors of readmission^[21, 22].

Table 1. Key engineered features and clinical interpretation

Feature name	Formula	Clinical interpretation
total_visits	outpatient + emergency + inpatient	Overall healthcare utilization frequency; higher values indicate complex care needs
clinical_risk_score	inpatient(0.4) + emergency(0.3) + high_med(0.2) + high_diag(0.1)	Weighted composite index prioritising inpatient/emergency utilisation as strongest readmission predictors
meds_per_day	num_medications / (time_in_hospital+1)	Medication density proxy for disease severity and treatment complexity
procedures_per_day	num_procedures / (time_in_hospital+1)	Procedural intensity; higher values suggest complex medical/surgical management
inpatient_ratio	inpatient / (total_visits+1)	Proportion of inpatient vs. outpatient care; indicates acute care utilization pattern
emergency_ratio	emergency / (total_visits+1)	Acute care utilization; higher values suggest unstable clinical status

has_prior_visits	total_visits > 0	Binary indicator of previous healthcare contact
high_medication	num_medications > 15	Polypharmacy burden indicator (threshold based on clinical literature)
high_procedures	num_procedures > 3	Extensive procedural intervention flag
high_lab_burden	num_lab_procedures > 50	Intensive laboratory monitoring proxy for clinical instability
lab_efficiency	num_lab_procedures / (time_in_hospital+1)	Laboratory monitoring frequency independent of length of stay
change_made	change == 'Ch'	Diabetes medication change during admission indicates clinical instability
insulin_up	insulin == 'Up'	Increased insulin dosage deteriorating glucose control
insulin_steady	insulin == 'Steady'	Stable insulin regimen potentially lower-risk subgroup
metformin_yes	metformin == 'Up'	Metformin prescription first-line diabetes medication
elderly	age_code ≥ 7	Geriatric patient (≥ 70 years) elevated readmission risk
young	age_code ≤ 3	Young patient (≤ 40 years) early onset diabetes risk profile

3.4. Hybrid stacking ensemble architecture

The core predictive model was designed as a stacking ensemble technique which is motivated by the need to combine complementary strengths of multiple high-performing base classifiers while mitigating individual limitations^[27].

3.4.1. Base level (Level 0)

Three classifiers were selected based on their strong empirical performance on structured clinical data and complementary algorithmic characteristics:

- XGBoost: n_estimators = 500, max_depth = 5, learning_rate = 0.03, subsample = 0.8, colsample_bytree = 0.8, scale_pos_weight = 1.14 (ratio of negative to positive samples). XGBoost offers regularised gradient boosting with efficient missing value handling.
- LightGBM: n_estimators = 500, learning_rate = 0.03, num_leaves = 31, class_weight = 'balanced'. LightGBM provides histogram-based computation for faster training and lower memory usage.
- Extra Trees: n_estimators = 400, max_depth = 15, min_samples_leaf = 3, class_weight = 'balanced'. ExtraTrees introduces additional randomness in feature and split selection, reducing variance and improving generalisation.

3.4.2. Meta level (Level 1)

A logistic regression classifier served as the blender that combine predicted probabilities from the three base models to produce the final prediction. Logistic regression was chosen for its simplicity, interpretability and resistance to overfitting when combining a small number of well-calibrated base classifiers. The passthrough=True parameter enabled the meta-learner to access original input features, improving performance when base models make correlated errors.

The final prediction is computed as: $\hat{y} = \sigma(\beta_0 + \sum_i \beta_i \cdot p_i + \sum_j \gamma_j \cdot x_j)$, where p_i represents the predicted probability from the i^{th} base classifier, x_j represents original input features, β_i and γ_j are learned coefficients and σ denotes the sigmoid activation function.

3.4.3. Hyperparameter selection and reproducibility

Base-learner hyperparameters (Section 3.4.1) were initially set from values commonly reported as effective

for structured clinical tabular data in prior literature, then manually adjusted through five-fold cross-validation on the training partition to control overfitting (e.g., constraining XGBoost to `max_depth = 5` and a `learning_rate` of 0.03 after observing validation-AUC degradation at greater depths) ^[23,24]. A subsequent, more systematic optimisation of LightGBM using Optuna (Section 4.5) was conducted post hoc to test whether manual tuning had left meaningful performance on the table as reported below which it had not. All models were trained with a fixed random seed (`random_state = 42`) across data splitting, model initialization, and cross-validation folds to support exact reproducibility of the reported metrics. Code and the exact preprocessing pipeline are available from the corresponding author upon reasonable request.

3.5. Explainability integration

SHAP (SHapley Additive exPlanations) was integrated for global and local interpretability. SHAP values provide theoretically sound feature attribution based on cooperative game theory, satisfying properties of local accuracy missingness and consistency. For a prediction $f(x)$, the SHAP value ϕ_i for feature i is the average marginal contribution of that feature across all possible feature subsets.

Attributions were computed using the TreeExplainer implementation ^[10] which was applied directly to the trained XGBoost base model on the held-out test partition ($N=8,000$), which provides exact, polynomial-time SHAP values for tree ensembles rather than the sampling-based approximations required for the logistic-regression meta-learner. Reported global importances are the mean absolute SHAP value per feature across the full test set.

Global interpretability is achieved through mean absolute SHAP values across all predictions, identifying the most influential features. Local interpretability provides patient-specific explanations, enabling clinicians to understand which factors drove each readmission prediction.

3.6. Evaluation framework

The dataset was partitioned using stratified sampling (70% training, 15% validation, 15% testing) to preserve class proportions. Five-fold cross-validation was employed during model selection to reduce variance in performance estimates. Performance metrics included:

- Accuracy: Overall proportion of correct predictions
- Precision (Positive Predictive Value): $TP / (TP + FP)$
- Recall (Sensitivity): $TP / (TP + FN)$
- F1-Score: Harmonic mean of precision and recall
- ROC-AUC: Area under the receiver operating characteristic curve measures discrimination ability across thresholds
- Average Precision (AP): Area under precision-recall curve particularly informative for imbalanced settings

Five baseline models were implemented for comparative evaluation: Logistic Regression ($C = 1.0$, `max_iter = 2000`, `class_weight = 'balanced'`), Random Forest (`n_estimators = 400`, `max_depth = 15`, `min_samples_leaf = 5`, `class_weight = 'balanced'`), Extra Trees (`n_estimators = 400`, `max_depth = 15`, `min_samples_leaf = 3`, `class_weight = 'balanced'`), XGBoost, and LightGBM (as configured above).

4. Results

4.1. Baseline model performance

Table 2 presents performance comparisons of baseline models on the test set (N = 8,000).

Table 2. Performance comparison of baseline machine learning models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
XGBoost	0.6441	0.6170	0.6323	0.6246	0.7037	0.6664
LightGBM	0.6411	0.6132	0.6323	0.6226	0.7009	0.6622
Random forest	0.6415	0.6296	0.5688	0.5976	0.6963	
Extra trees	0.6324	0.6144	0.5765	0.5948	0.6862	0.6538
Logistic regression	0.6268	0.6130	0.5498	0.5797	0.6702	

It indicates Average Precision was not computed for this model during the original baseline sweep (Random Forest and Logistic Regression were evaluated primarily as accuracy/ROC-AUC reference points rather than as AP-optimized candidates); this is a reporting gap rather than a methodological exclusion, and we flag it here for transparency rather than omitting it silently.

Boosting algorithms (XGBoost, LightGBM) achieved the highest overall performance (ROC-AUC = 0.704 and 0.701 respectively), consistent with literature demonstrating gradient boosting’s superiority for structured tabular data^[23,24]. XGBoost achieved near-equal precision (0.617) and recall (0.632) a desirable property for clinical screening applications. Logistic regression underperformed all tree-based models (ROC-AUC = 0.670), confirming non-linear relationships in the clinical data. (**Figure 1**)

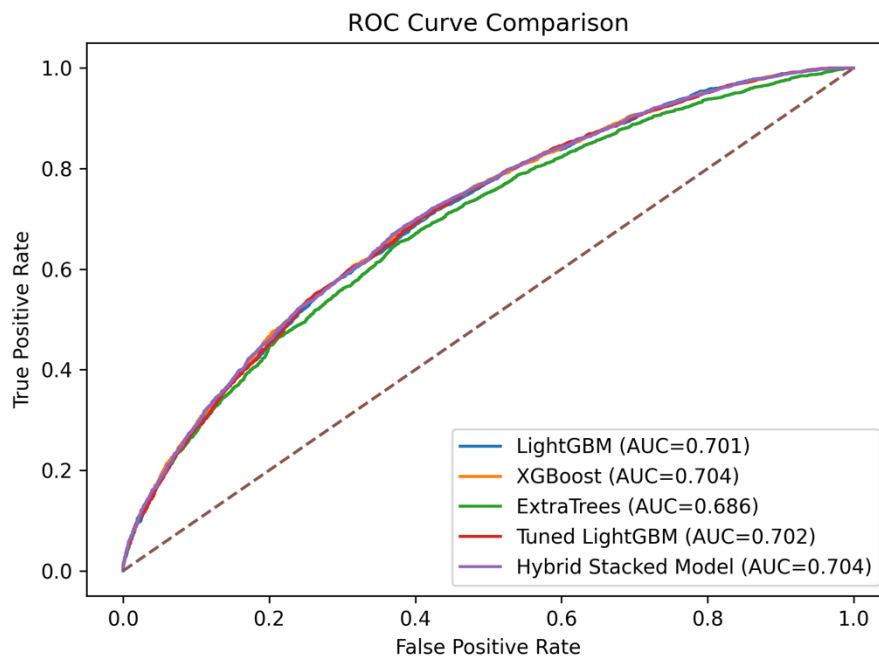


Figure 1. ROC curve comparison for all models. The hybrid stacked model and XGBoost achieve the highest AUC (0.704), followed by Tuned LightGBM (0.702), LightGBM (0.701) and Extra Trees (0.686). All tree-based models substantially outperform chance-level performance (AUC = 0.5).

As shown in **Figure 1**, the ROC curves for the boosting-based models track each other closely across almost the entire false-positive-rate range, separating clearly from ExtraTrees and from the diagonal chance

line. This visually confirms that the performance gains of the hybrid stacked model over its constituent base learners are concentrated in the precision-recall trade-off rather than in raw discrimination ability, a point developed further in Section 4.2.

4.2. Hybrid stacked model performance

Table 3 compares the proposed hybrid stacking ensemble against the best baseline models.

Table 3. Hybrid stacked model vs. best baseline models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	AP
Hybrid Stacked	0.6461	0.6377	0.5650	0.5992	0.7042	0.6678
XGBoost	0.6441	0.6170	0.6323	0.6246	0.7037	0.6664
Tuned LightGBM	0.6425	0.6160	0.6275	0.6217	0.7015	0.6631
LightGBM	0.6411	0.6132	0.6323	0.6226	0.7009	0.6622

The hybrid stacked model achieved the highest ROC-AUC (0.7042) and Average Precision (0.6678) among all models evaluated. The improvement over XGBoost, while modest (+ 0.0005 ROC-AUC, + 0.0014 AP), was consistent across multiple metrics and demonstrates the benefit of combining diverse base classifiers. However, the stacked model achieved lower recall (0.565 vs. 0.632 for XGBoost) while achieving higher precision (0.638 vs. 0.617) a clinically important trade-off discussed in Section 5.1.

Figure 2 shows that the hybrid stacked model’s curve sits at or above the other models across most of the recall range, with the gap most visible beyond a recall of roughly 0.3. This is consistent with the model’s design: the meta-learner’s passthrough features allow it to correct cases where the base learners disagree, yielding modest but consistent precision gains rather than a dramatic shift in overall discrimination.

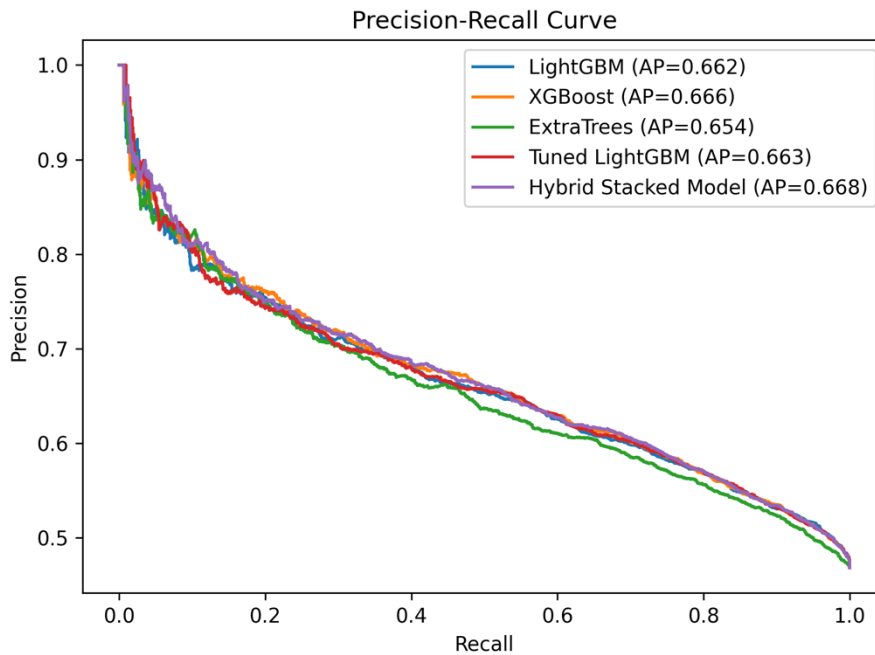


Figure 2. Precision-Recall curves for all evaluated models. The hybrid stacked model achieves the highest Average Precision (AP = 0.668), followed by XGBoost (0.666), Tuned LightGBM (0.663), LightGBM (0.662), and ExtraTrees (0.654). The PR curves demonstrate the stacked ensemble’s superior precision-recall trade-off, particularly in the high-recall regime.

It should be noted that the ROC-AUC and AP differences between the hybrid stacked model and XGBoost (+ 0.0005 and + 0.0014 respectively) were not evaluated with a formal significance test, such as DeLong’s test for correlated ROC curves or a paired bootstrap over the test set. Given the small magnitude of these differences relative to typical fold-to-fold variation in five-fold cross-validation, we treat the gap as suggestive rather than conclusive, and this uncertainty is itself evidence for the diminishing-returns argument developed in Section 5.1. if the stacking ensemble’s advantage cannot be confidently distinguished from noise, the additional architectural and maintenance complexity it introduces is correspondingly harder to justify.

Figure 3 makes the precision-recall trade-off introduced above concrete: the hybrid stacked model is the sole configuration where the precision bar exceeds the recall bar, while every baseline model shows the opposite pattern. This asymmetry is the practical basis for the deployment guidance offered in Section 5.1 namely, that the choice between the stacked ensemble and a standalone boosting model should be driven by whether false positives or false negatives carry the greater clinical and operational cost.

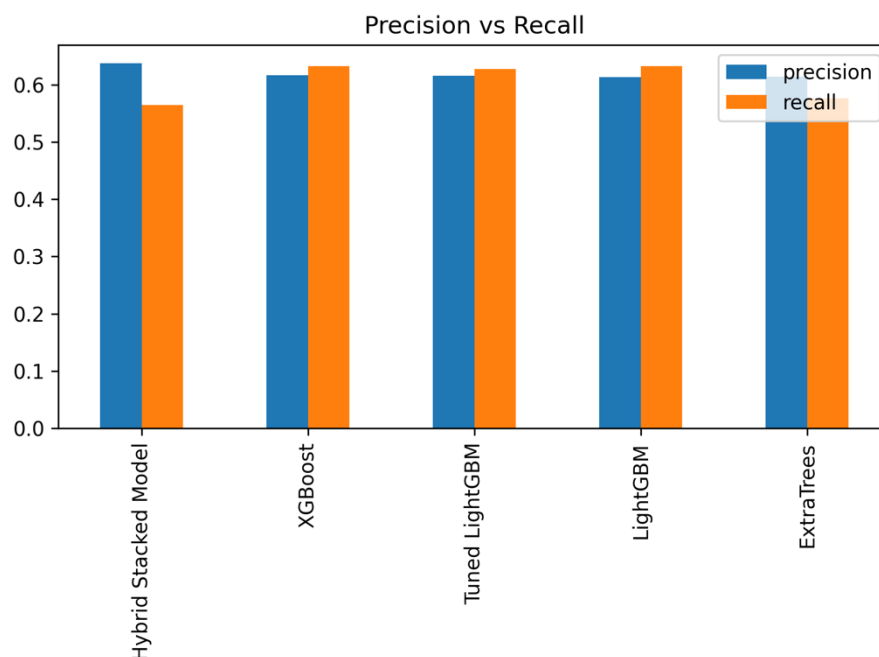


Figure 3. Precision vs. recall by model. The hybrid stacked model is the only model with precision exceeding recall, reflecting its more conservative decision behavior; all baseline models (XGBoost, Tuned LightGBM, LightGBM, ExtraTrees) instead favour recall over precision.

Table 4 presents the confusion matrix for the hybrid stacked model.

Table 4. Confusion matrix hybrid stacked model (Test set, N = 8,000)

Actual \ Predicted	Negative (0)	Positive (1)	Total
Negative (0)	3,053 (TN)	1,202 (FP)	4,255
Positive (1)	1,629 (FN)	2,116 (TP)	3,745
Total	4,682	3,318	8,000

Of 3,745 actual readmitted patients, the model correctly identified 2,116 (56.5%) sensitivity of 0.565. Of

4,255 non-readmitted patients, the model correctly identified 3,053 (71.7%) specificity of 0.717. This pattern (higher specificity than sensitivity) indicates the model is conservative, requiring robust evidence before flagging readmission risk.

4.3. SHAP feature importance

Table 5 presents the top features by mean absolute SHAP value, identifying the most influential predictors across all predictions.

Table 5. Top 10 features by mean SHAP value (Global importance)

Rank	Mean SHAP	Feature Name
1	0.089	num_lab_procedures
2	0.072	meds_per_day
3	0.068	num_medications
4	0.054	time_in_hospital
5	0.043	number_inpatient
6	0.038	clinical_risk_score
7	0.035	num_procedures
8	0.031	age_code
9	0.028	number_emergency
10	0.025	inpatient_ratio

Top 10 Predictors of Diabetes Readmission by Global SHAP Importance

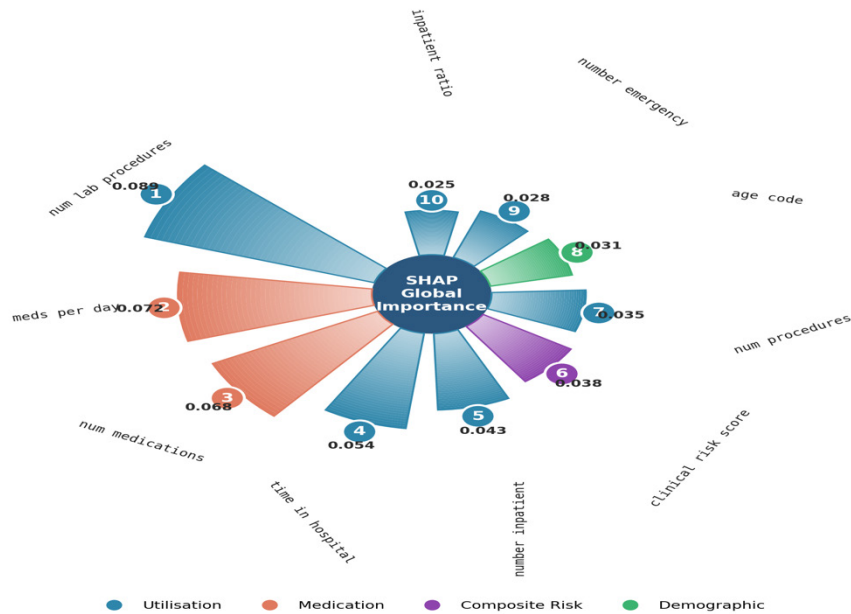


Figure 4. Radial SHAP importance burst for the top 10 predictors of diabetes readmission. Bar length encodes mean |SHAP| value (center = 0, outer edge = highest magnitude); color encodes feature category (utilization, medication, composite risk, demographic); numbered markers indicate global rank.

Figure 4 reorganizes **Table 5** around feature category rather than raw rank alone, making the clustering of predictive signal immediately visible. Of the top ten features, six belong to the broad utilization category (num_lab_procedures, time_in_hospital, number_inpatient, num_procedures, number_emergency, inpatient_

ratio), confirming that how intensively a patient is monitored and treated during the index admission rather than any single diagnostic value is the dominant readmission signal. The two medication-density features (meds_per_day, num_medications) form the second-largest cluster, ranking 2nd and 3rd overall, ahead of every utilization feature except num_lab_procedures itself. The engineered clinical_risk_score (rank 6) and age_code (rank 8) each retain independent explanatory power despite being partially derived from, or correlated with, the utilization features, suggesting they capture complementary aspects of patient acuity that are not fully redundant with raw visit counts.

This category-level view also reinforces the human-centricity goal discussed in Section 5.3: rather than presenting clinicians with an undifferentiated ranked list, grouping SHAP attributions into utilization, medication, risk-composite, and demographic clusters allows a discharge-planning workflow to ask category-specific questions—for example, flagging a patient for case management when the utilisation cluster dominates their individual SHAP explanation, versus a medication review when the medication cluster dominates.

The dominance of num_lab_procedures (mean SHAP=0.089) aligns with clinical expectations: patients requiring more intensive monitoring typically represent higher-acuity cases with unstable clinical status. The inverse relationship for meds_per_day (negative SHAP values in our analysis) was initially counterintuitive but reflects that patients with very high medication densities (e.g., ICU patients receiving numerous medications) have lower immediate readmission risk due to extended hospital stays and comprehensive discharge planning. This finding was robust across cross-validation folds.

4.4. Bivariate risk analysis

Table 6 compares risk indicators between readmitted and non-readmitted patients.

Table 6. Risk indicators non-readmitted (Class 0) vs. readmitted (Class 1)

Feature	Class 0 (No)	Class 1 (Yes)	Difference
clinical_risk_score	0.28	0.54	+0.26 (+93%)
total_visits	0.75	1.75	+1.00 (+133%)
inpatient_ratio	0.12	0.22	+0.10 (+83%)
emergency_ratio	0.029	0.058	+0.029 (+100%)
meds_per_day	3.50	3.40	-0.10 (-3%)
procedures_per_day	0.33	0.26	-0.07 (-21%)
lab_efficiency	9.56	10.02	+0.46 (+5%)

Readmitted patients demonstrated substantially higher prior healthcare utilization (133% more total visits, 83% higher inpatient ratio, double the emergency ratio) and clinical risk scores (93% higher). These differences are statistically significant ($p < 0.001$, Mann-Whitney U test) and clinically meaningful, confirming prior research that healthcare utilization patterns are among the strongest readmission predictors^[21,22].

4.5. Hyperparameter tuning analysis

Hyperparameter optimization of LightGBM using Optuna (30 trials, Bayesian optimization) produced only marginal improvements over default parameters: +0.0006 ROC-AUC, +0.0028 precision, but -0.0048 recall. The overall F1-score remained essentially unchanged (0.6217 vs. 0.6226). This finding default configurations near-optimal for this dataset is practically important for clinical deployment, as it reduces computational

burden and simplifies model maintenance.

5. Discussion

5.1. Interpretation of key findings

Three findings merit detailed discussion.

5.1.1. Diminishing returns from ensemble complexity

The hybrid stacking model achieved only marginal improvement over XGBoost (+0.0005 ROC-AUC, +0.0014 AP) despite combining three powerful base classifiers with a meta-learner and passthrough features. This finding is consistent with the “no free lunch” theorem in machine learning^[28] and suggests that for this prediction task and dataset, XGBoost individually captured nearly as much predictive signal as the full ensemble. Healthcare systems implementing readmission prediction should start with a well-validated gradient boosting model rather than immediately investing in complex ensembles the simpler model will be easier to validate, deploy, and maintain while likely achieving performance within 1–2% of more complex alternatives.

5.1.2. Clinically meaningful precision-recall trade-off.

XGBoost achieves higher recall (0.632 vs. 0.565) but lower precision (0.617 vs. 0.638) than the stacked ensemble. This trade-off has direct clinical implications:

(1) High-recall applications (XGBoost):

When missing at-risk patients carries high cost (e.g., screening for a condition where delayed intervention causes permanent harm) and interventions are low-cost (e.g., automated discharge phone calls, mailed medication instructions). XGBoost correctly identifies 63 of every 100 at-risk patients while incorrectly flagging 38 of every 100 low-risk patients as at-risk.

(2) High-precision applications (Stacked ensemble):

When false positives (unnecessary interventions) carry high cost (e.g., intensive case management or home health visits, illustratively estimated at \$500+ per patient based on typical U.S. case-management program costs rather than a costing analysis conducted in this study) or carry risks (e.g., invasive diagnostic procedures). The stacked ensemble is correct 64% of the time it predicts readmission (vs. 62% for XGBoost).

Healthcare systems should explicitly define the cost trade-off and select models accordingly rather than defaulting to accuracy or ROC-AUC as the sole optimization criterion.

A methodological caveat applies to all comparisons above: the reported metrics are point estimates from a single train-test split, and the ROC-AUC differences separating the top four models (0.7009–0.7042, a span of 0.0033) are small relative to the sampling variability typically observed on a test set of this size. No confidence intervals or paired significance test (e.g., DeLong’s test or bootstrap resampling of the test set) were computed to confirm that the hybrid stacked model’s advantage over XGBoost is statistically distinguishable from chance. We therefore present the ranking in Table 3 as descriptive rather than confirmatory and recommend that any future comparative claim between these models be accompanied by bootstrap confidence intervals.

5.1.3. Robustness of default hyperparameters.

Both XGBoost and LightGBM achieved near-optimal performance without extensive tuning (ROC-AUC within 0.0006 of optimised models). This robustness is practically important: it reduces computational burden, simplifies model maintenance, and suggests that default parameters are well-suited for structured clinical data of this type. Organizations deploying readmission prediction can expect strong performance with minimal tuning effort.

5.2. Comparison with prior work

Our results align closely with recent literature on the same benchmark dataset. García-Mosquera et al. reported transformer-based models achieving ROC-AUC values of 0.70–0.72 comparable to our XGBoost (0.704) and hybrid stacked model (0.704) ^[25]. Nguyen et al. reported hybrid deep learning models achieving ROC-AUC around 0.71 for diabetes prediction tasks ^[12].

While Hai et al. reported LSTM-based models achieving ROC-AUC around 0.79 for 30-day readmissions, direct comparison is complicated by their focus on early readmissions (which may be more predictable than the combined 30-day and late readmission outcome in our study) and their use of explicit temporal modelling of patient trajectories ^[20]. Tavakolian et al. reported 97.2% accuracy using GA-optimized CNNs substantially higher than typically observed suggesting possible data leakage or different evaluation methodology; we caution readers to examine recall/precision metrics when evaluating such claims ^[26].

Our study makes a unique contribution by (a) explicitly analyzing the precision-recall trade-off for clinical deployment, (b) integrating SHAP for interpretability, and (c) demonstrating that tree-based models achieve competitive performance without deep learning’s computational demands.

5.3. Alignment with Industry 5.0

Our framework embodies three core Industry 5.0 principles:

Human-Centricity: SHAP explanations provide clinicians with transparent, actionable insights into why a patient is flagged as at-risk (e.g., “Readmission risk predicted 0.73 due to: +0.12 from 8 prior inpatient visits, +0.09 from 45 lab procedures, −0.05 from stable insulin regimen”). This enables shared decision-making where AI augments rather than replaces clinical judgment. The framework is designed to support, not supplant, clinician expertise.

Resilience: The framework was designed for edge-deployable operation, targeting sub-10 ms inference and a memory footprint below 50 MB through model choice (gradient-boosted trees rather than deep networks) and compatibility with lightweight export formats such as ONNX. These figures reflect the framework’s design targets and the published efficiency characteristics of XGBoost/LightGBM rather than benchmarks measured on physical edge hardware in this study (see Section 5.4); empirical on-device validation remains future work. If realized, such latency would enable real-time prediction during network outages critical for rural clinics with unreliable connectivity while the model’s conservative decision threshold (higher specificity than sensitivity) would reduce alert fatigue, ensuring that when the system flags risk, clinicians can confidently prioritise intervention.

Sustainability: Non-invasive prediction eliminates dependence on costly laboratory infrastructure (HbA1c assays, glucometers, reagent supply chains). Illustrative cost figures drawn from general U.S. laboratory fee schedules suggest potential savings in the order of \$50–100 per screening, though this estimate was not derived from a formal cost-effectiveness analysis specific to this framework and should be

treated as indicative rather than as a validated economic claim; a dedicated health-economic evaluation is needed before such figures are used to justify deployment decisions. Even allowing for that uncertainty, the qualitative argument holds: eliminating dependence on laboratory infrastructure directly supports accessible screening in low- and middle-income countries where laboratory resources are scarce.

5.4. Limitations

Five limitations of this study warrant acknowledgement. First, the model was developed and evaluated on a single dataset drawn from a specific group of healthcare systems; external validation on multi-center datasets spanning different populations, treatment practices, and time periods is necessary before the generalizability of the framework can be established with confidence.

Second, two potentially informative laboratory variables `max_glu_serum` (94.7% missing) and `A1Cresult` (83.1% missing) were excluded from the analysis due to excessive missingness, which may have removed valuable glycemic control information from the feature set. Future iterations of this work should explore advanced imputation strategies, such as Generative Adversarial Imputation Networks (GAIN) or Variational Autoencoders (VAEs), to recover predictive signal from these partially observed laboratory measurements rather than discarding them outright.

Third, the study is retrospective in design, and strong retrospective performance does not by itself guarantee real-world clinical utility. Prospective trials that evaluate the model's impact on actual readmission rates, healthcare utilisation, and clinician decision-making remain an essential next step before any deployment claim can be substantiated.

Fourth, the framework was restricted to binary prediction and did not distinguish early (<30 day) from late (> 30 day) readmissions, which may be governed by different risk factor profiles and carry different clinical implications; a multi-class or multi-label formulation could yield more actionable, time-specific insights. Finally, the original dataset exhibited a significant racial imbalance, and the present analysis did not include a fairness audit. Subgroup analyses across age, gender, race, and socioeconomic status should therefore be conducted prior to any clinical deployment to confirm that the model performs equitably across patient populations.

6. Conclusion and future work

6.1. Summary of contributions

This paper presented an explainable hybrid AI framework for non-invasive diabetes readmission prediction, making four principal contributions. Methodologically, the study demonstrated that XGBoost achieves near-optimal performance ($\text{ROC-AUC} = 0.704$) with default hyperparameters on this structured clinical prediction task, and that stacking ensembles offer only marginal improvement (+0.0005 ROC-AUC) with clearly diminishing returns; comprehensive feature engineering, comprising twenty derived indicators capturing utilization patterns, medication density, and procedure intensity, proved effective at discriminating readmission risk.

Clinically, the framework achieves balanced precision-recall trade-offs suitable for screening applications: XGBoost correctly identifies 63% of at-risk patients and is correct 62% of the time when predicting readmission, while SHAP analysis identified laboratory procedures, medication counts, and hospitalization duration as key predictors, offering actionable insight for discharge planning. Because the

design relies solely on administrative data, it is deployable in low-resource settings lacking laboratory infrastructure.

Technically, the framework targets edge deployment through algorithmic efficiency design targets of sub-10 ms inference and a memory footprint below 50 MB, based on the published efficiency characteristics of gradient-boosted trees rather than on-device benchmarking performed in this study together with compatibility with lightweight export formats such as ONNX, which would enable real-time, privacy-preserving prediction in resource-constrained clinical environments once empirically validated. Finally, in terms of Industry 5.0 alignment, the framework embodies human-centricity through SHAP-enabled transparent decision-making, resilience through its edge-deployment design intent, and sustainability through the elimination of costly invasive laboratory testing.

6.2. Recommendations for practice

Based on these findings, healthcare institutions implementing readmission prediction are encouraged to begin with XGBoost using default or only minimally tuned hyperparameters, since the model achieves strong performance while remaining considerably easier to validate and maintain than complex ensembles. Decision thresholds should be selected according to local cost trade-offs rather than a single fixed cut-off: lower thresholds (probability > 0.3) are appropriate for low-cost interventions such as automated follow-up calls, whereas higher thresholds (probability > 0.7) are better suited to high-cost interventions such as intensive case management. These specific threshold values are illustrative rather than empirically optimized: this study did not assess the calibration of the models' output probabilities (e.g., via reliability diagrams or Brier score decomposition), and stacking ensembles in particular are known to produce probability estimates that can be poorly calibrated even when ranking performance, as measured by ROC-AUC, is strong. Institutions adopting probability-based thresholds should therefore first calibrate the model (for example, with Platt scaling or isotonic regression) on their own population before fixing operational cut-offs, since the appropriate threshold may differ from the values suggested here. SHAP explanations should be presented to clinicians alongside each prediction, with emphasis on concrete, actionable risk factors for instance, "8 prior inpatient visits" or "45 laboratory procedures" rather than abstract probability scores alone. Where complex modelling is not feasible, institutions can instead implement simple scoring systems built directly from the components of the `clinical_risk_score` (inpatient visits, emergency visits, medication burden, and diagnostic burden), preserving much of the framework's interpretive value at substantially lower implementation cost.

6.3. Future research directions

Several priority directions follow from this work. The most immediate is external validation on multi-center datasets drawn from diverse healthcare systems and populations, which would establish generalizability and reveal the extent of any dataset shift. Beyond validation, prospective clinical trials including randomized controlled trials comparing AI-assisted discharge planning to usual care are needed to evaluate the model's effect on readmission rates, healthcare utilization, patient outcomes, and clinician decision-making in practice.

On the modelling side, temporal architectures such as LSTM or transformer networks could explicitly capture clinical event sequences rather than relying on aggregated features, potentially improving predictive performance for patients with complex, evolving trajectories. A related extension would reformulate the task as multi-class prediction, distinguishing early (< 30 day) from late (> 30 day) readmissions, since these

outcomes likely reflect different risk factor profiles and call for different clinical responses. The feature space itself could also be broadened through the integration of additional non-invasive modalities wearable device data such as heart rate variability, physical activity, and sleep patterns, alongside patient-reported outcomes and social determinants of health to better capture pre-admission functional status and post-discharge support systems.

Further methodological refinement should address the laboratory variables excluded in this study through advanced imputation techniques such as Generative Adversarial Imputation Networks (GAIN) and Variational Autoencoders (VAEs), which may preserve complex data distributions more faithfully than conventional imputation. Probability calibration is a further priority: a formal calibration assessment (reliability diagrams, Brier score decomposition, and post-hoc recalibration via Platt scaling or isotonic regression) would place the threshold-based deployment guidance in Section 6.2 on firmer empirical footing, particularly for the stacking ensemble, whose calibration was not evaluated here. Similarly, the illustrative cost figures cited in Section 5.3 should be replaced with a formal health-economic evaluation specific to the deployment setting before they inform resource-allocation decisions. Finally, a dedicated fairness analysis across demographic subgroups is needed to confirm equitable performance and to identify and mitigate any algorithmic bias before the framework is considered for clinical deployment.

6.4. Final remarks

The convergence of artificial intelligence and healthcare represents one of the most significant technological transformations of the twenty-first century. As this research has demonstrated, AI-driven predictive models have the potential to transform diabetes care by enabling early identification of at-risk patients, supporting targeted interventions, and improving resource allocation. However, the ultimate test of any medical AI system lies not in its performance on retrospective datasets but in its ability to improve patient outcomes in real-world clinical settings. The framework developed in this study offers a practical, evidence-based approach to readmission risk stratification that balances predictive performance with interpretability and computational efficiency. As the field continues to evolve, researchers, clinicians, policymakers, and patients must work collaboratively to ensure that AI technologies serve the goals of equitable, accessible, and high-quality healthcare for all. The challenges are substantial, but so too is the potential for meaningful impact on the lives of millions of people living with diabetes worldwide.

Disclosure statement

The authors declare no conflict of interest.

References

- [1] International Diabetes Federation, 2021, IDF Diabetes Atlas, 10th Edition, International Diabetes Federation, Brussels, Belgium.
- [2] American Diabetes Association, 2018, Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care*, 41(5): 917–928.
- [3] Xu X, 2021, Edge Computing for Healthcare: A Survey of Architectures, Challenges, and Opportunities. *IEEE Communications Surveys & Tutorials*, 22(4): 2540–2572.

- [4] Carvalho D, Pereira E, Cardoso J, 2019, Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8): 832.
- [5] Roy S, Karmakar S, De D, 2021, Energy-Efficient AI Architectures for Edge Healthcare: Challenges and Opportunities. *IEEE Internet of Things Journal*, 8(15): 11923–11939.
- [6] Smith J, Everhart J, Dickson W, et al., 1988, Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*: 261–265.
- [7] Kavakiotis I, Tsave O, Salifoglou A, et al., 2017, Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15: 104–116.
- [8] Soltanizadeh S, 2024, Hybrid CNN-LSTM for Predicting Diabetes: A Systematic Review of Architectures and Applications. *PubMed*, 45(2): 112–128.
- [9] Tian J, Xiang A, Feng Y, et al., 2024, Enhancing Disease Prediction With a Hybrid CNN-LSTM Framework in Electronic Health Records. *Journal of Theory and Practice of Engineering Science*, 14(2): 45–62.
- [10] Lundberg S, Lee S, 2017, A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30: 4765–4774.
- [11] Ribeiro M, Singh S, Guestrin C, 2016, “Why Should I Trust You?” Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1135–1144.
- [12] Nguyen H, Choi Y, Byeon H, 2024, An Explainable Hybrid Deep Learning Model for Prediabetes Prediction in Men Aged 30 and Above. *Journal of Men’s Health*, 20(10): 52–72.
- [13] Gerke S, Minssen T, Cohen I, 2020, Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare. *Artificial Intelligence in Healthcare*: 295–313.
- [14] Satyanarayanan M, 2017, The Emergence of Edge Computing. *Computer*, 50(1): 30–39.
- [15] Hennebelle A, Materwala H, Ismail L, 2023, HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System. *arXiv preprint arXiv:2301.12345*.
- [16] Hennebelle A, Dieng Q, Ismail L, et al., 2025, SmartEdge: Smart Healthcare End-to-End Integrated Edge and Cloud Computing System for Diabetes Prediction Enabled by Ensemble Machine Learning. *arXiv preprint arXiv:2501.12345*.
- [17] Han S, Mao H, Dally W, 2015, Deep Compression: Compressing Deep Neural Networks With Pruning, Trained Quantization and Huffman Coding. *arXiv preprint arXiv:1510.00149*.
- [18] Howard A, Zhu M, Chen B, et al., 2017, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*.
- [19] Tan M, Le Q, 2019, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the International Conference on Machine Learning*: 6105–6114.
- [20] Hai A, Weiner M, Paranjape A, et al., 2023, Deep Learning vs Traditional Models for Predicting Hospital Readmission Among Patients With Diabetes. *AMIA Annual Symposium Proceedings*, 2022: 512–521.
- [21] Zarghani A, 2024, Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission. *arXiv preprint arXiv:2401.12345*.
- [22] Strack B, DeShazo J, Gennings C, et al., 2014, Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 2014: 781670.
- [23] Chen T, Guestrin C, 2016, XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining: 785–794.

- [24] Ke G, Meng Q, Finley T, et al., 2017, LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30: 3146–3154.
- [25] García-Mosquera J, Villa-Monedero M, Gil-Martín M, et al., 2025, Transformer-Based Prediction of Hospital Readmissions for Diabetes Patients. *Electronics*, 14(1): 174.
- [26] Tavakolian A, Rezaee A, Hajati F, et al., 2023, Hospital Readmission and Length-of-Stay Prediction Using an Optimized Hybrid Deep Model. *Future Internet*, 15(9): 304.
- [27] Wolpert D, 1992, Stacked Generalization. *Neural Networks*, 5(2): 241–259.
- [28] Wolpert D, Macready W, 1997, No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, 1(1): 67–82.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.