Beyond the Black Box: Clinically Engineered Features and a Composite Risk Score for Diabetes Readmission Prediction

Aminu Usman Jibril
Department of Computer Science, Skyline University Nigeria, Kano, Kano State, Nigeria

A. Senthil Kumar
Dean, School of Science and Information Technology (SSIT), Department of Computer Science, Skyline University Nigeria

Abdullahi Shehu Ahmad
Computer Science Department, Faculty of Computing, Bayero University, Kano, Kano State, Nigeria

ABSTRACT
Hospital readmission among patients with diabetes remains an important clinical and economic challenge, yet predictive-modelling research often emphasises algorithmic solutions over clinically grounded feature derivation. This study developed and evaluated a systematic, interpretable feature-engineering framework for stratifying diabetes readmission risk using non-invasive administrative and clinical data extracted from electronic health records. The analysis used 40,000 encounters from the Diabetes 130-US Hospitals dataset. Twenty engineered predictors were derived across seven domains: visit history, medication burden, procedural utilisation, laboratory utilisation, composite clinical risk, medication-change patterns, and age-related indicators; diagnosis codes were mapped into 16 clinically coherent categories. A hybrid stacking ensemble combining XGBoost, LightGBM and Extra Trees with a logistic-regression meta-learner was benchmarked against baseline classifiers. Performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC and average precision (AP), while SHAP was used to interpret feature contributions. The hybrid stacking model achieved the highest ROC-AUC (0.7042), AP (0.6678), accuracy (0.6461) and precision (0.6377), whereas XGBoost achieved higher recall (0.6323) and F1-score (0.6246). The number of laboratory procedures, medication burden relative to length of stay (meds_per_day), total medication count, hospital duration, diagnostic categories and prior inpatient utilisation emerged as important predictors. The composite clinical_risk_score was higher among readmitted than non-readmitted encounters (mean 0.54 vs. 0.28), supporting its value as an interpretable risk representation. These findings indicate that clinically informed feature engineering can yield actionable and explainable predictors of diabetes readmission while showing that added ensemble complexity provides only marginal discrimination gains. External validation and prospective evaluation remain necessary before clinical deployment.

General Terms
Machine Learning, Healthcare Informatics, Clinical Decision Support, Explainable Artificial Intelligence.

Keywords
Diabetes readmission, feature engineering, clinical risk score, explainable artificial intelligence, electronic health records, machine learning, risk stratification.

1. INTRODUCTION
Diabetes mellitus is a major contributor to unplanned hospital utilisation and imposes substantial clinical and financial pressure on healthcare systems [1]. Diabetes-related readmissions remain persistently high, reflecting the chronic multisystem nature of the disease and continuing challenges in discharge planning and post-acute care coordination [2]. In the United States, the economic burden of diagnosed diabetes was estimated at approximately $327 billion in 2017, with inpatient services representing a substantial component of this cost [1].

Most machine-learning research on diabetes readmission has focused on improving algorithmic performance, often applying ensemble or deep-learning architectures to raw EHR variables with comparatively less attention to how those variables are clinically transformed before modelling [3]. Gradient-boosting methods such as XGBoost [4] and LightGBM [5] have substantially advanced predictive modelling on structured data. However, models built directly from raw administrative fields may fail to capture clinically intuitive constructs such as medication burden relative to length of stay or composite illness complexity. This emphasis on algorithmic technique over feature-level clinical reasoning can limit interpretability and clinical adoption, particularly where laboratory and physiological data are incomplete [3].

Explainable artificial intelligence methods, particularly SHAP, provide a partial remedy by attributing model predictions to individual input features [6]. Nevertheless, post-hoc explanation is only as clinically meaningful as the underlying representation being explained. Explaining a prediction in terms of clinically opaque raw fields does not necessarily generate actionable insight. Limited evidence exists on how readily available EHR-derived utilisation patterns, medication history and demographic indicators can be transformed into a compact, clinically interpretable risk representation suitable for discharge-planning support.

This study addresses that gap by developing a clinically interpretable feature-engineering and risk-stratification framework for diabetes readmission prediction. Rather than relying on algorithmic complexity alone, the framework constructs 20 clinically motivated features across seven domains and combines selected utilisation and burden indicators into a composite clinical_risk_score. Five baseline classifiers and a hybrid stacking ensemble are evaluated, and SHAP is used to examine whether the most influential predictors remain clinically plausible.

2. PROBLEM STATEMENT
This study formalizes diabetes hospital readmission risk stratification as a supervised binary classification problem defined over a clinically engineered feature space. Framing the

task in this manner rather than as an unconstrained pattern-recognition task is scientifically justified for three reasons. Firstly, the clinical outcome of interest which is the readmission within the observation window is inherently dichotomous, making binary classification with natural statistical formalism. Secondly, expressing the problem as a regularized empirical risk minimization situates the analysis within a well-established statistical learning framework which allow principled comparison across heterogeneous model classes using a common objective. Thirdly, explicit mathematical notation makes every subsequent modelling choice from feature construction to ensemble combination that is traceable to a transparent falsifiable specification rather than an ad hoc procedure.

Let $D = \{(r_i, y_i)\}_{i=1}^{N}$ denote the observed dataset of N inpatient encounters where r_i is the raw electronic health record (EHR) for encounter i that comprising demographic, administrative, pharmacologic, procedurals and diagnostic fields and $y_i \in \{0,1\}$ is the binary readmission label with $y_i = 1$ indicating readmission and $y_i = 0$ indicating non-readmission. A feature engineering mapping

$$\varphi: \mathcal{R} \to \mathbb{R}^p \qquad (1)$$

transforms each raw record into a p-dimensional engineered feature vector $x_i = \varphi(r_i)$, and a predictive function

$$f_\theta: \mathbb{R}^p \to [0,1], \qquad \hat{p}_i = f_\theta(x_i) \approx P(y_i = 1 \mid x_i) \qquad (2)$$

parameterized by θ, estimates the posterior probability of readmission given the engineered features. The learning objective is the regularized empirical risk minimization problem

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f_\theta(x_i)\right) + \lambda\, \Omega(\theta) \qquad (3)$$

where $\mathcal{L}(y_i, \hat{p}_i)$ is the loss function penalizing disagreement between predicted and observed outcomes, $\Omega(\theta)$ is a model-complexity penalty, and $\lambda \geq 0$ controls the strength of regularization. Throughout this study, $\mathcal{L}$ is instantiated as the binary cross-entropy (log) loss,

$$\mathcal{L}(y_i, \hat{p}_i) = -[\, y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)\,] \qquad (4)$$

which is the maximum-likelihood-consistent loss for Bernoulli-distributed outcomes and penalizes confident incorrect predictions more heavily than uncertain ones a property that is clinically desirable given the asymmetric cost of missing a high-risk patient relative to mild miscalibration among low-risk patients.

The research is guided by three analytical hypotheses. H1 evaluates whether the clinically engineered representation improves discrimination relative to a raw-feature representation, with the target criterion AUC_engineered > AUC_raw. H2 evaluates whether the composite clinical risk score differs between outcome groups, μ_R|y=1 ≠ μ_R|y=0. H3 evaluates whether the hybrid stacking ensemble matches or exceeds the strongest individual learner in discrimination, AUC_stack ≥ max_m(AUC_m). The archived experimental outputs provide direct evidence for H2 and H3; a complete raw-only ablation required for formal inference on H1 was not retained, so no unsupported H1 effect estimate is introduced in this revision.

3. RELATED WORK

The clinical and economic challenge of diabetes-related hospital readmission has motivated two parallel research streams: epidemiological studies that identify discrete risk factors and machine-learning studies that build predictive models. Many studies, however, stop short of explicitly integrating these streams through systematic, clinically grounded feature engineering.

Foundational work by Strack et al. [2] established the Diabetes 130-US Hospitals corpus and demonstrated that HbA1c measurement patterns and healthcare-utilisation information were associated with readmission. Subsequent risk-factor studies refined this picture: Rubin et al. [7] identified treatment- and discharge-related predictors, while McDaniel and Chou [8] showed that social needs can add information beyond clinical variables alone. Together, these studies support the use of utilisation history, medication burden and broader clinical context, but they largely evaluate individual predictors rather than constructing explicit model-ready composite representations.

A second stream applies modern machine-learning methods directly to diabetes readmission data with comparatively limited feature transformation. Shang et al. [9] compared several classifiers on the same Health Facts-derived corpus, while Zarghani [10] benchmarked traditional machine-learning models against an LSTM architecture. Emi-Johnson and Nkrumah [11] evaluated logistic regression, random forest, XGBoost and deep learning on the same source dataset and identified prior admissions and medication count among important predictors. Mishra et al. [12] reported higher metrics in a much smaller cohort, illustrating why differences in sample size, outcome definition and validation design must be considered when comparing headline performance across studies.

A related line of work has pursued clinically scoped risk models rather than general-purpose classifiers. He et al. [13] developed and validated a 30-day readmission model for elderly patients with type 2 diabetes complicated by heart failure, using a compact set of clinically weighted predictors. This approach is conceptually aligned with the composite clinical_risk_score proposed here, although it targets a narrower comorbid population. More broadly, diabetes machine-learning reviews continue to identify interpretability and clinically grounded feature design as areas requiring further development [3], while SHAP-based explanation [6] addresses model transparency without replacing the need for meaningful input representation.

Finally, Mahmoudi et al. [14] reviewed EHR-based readmission risk-prediction models and reported substantial heterogeneity in feature selection and limited external validation. A recent systematic review of machine-learning approaches to hospital readmission similarly found inconsistent predictor curation across studies [15]. The present study addresses this convergent gap by treating feature engineering, rather than algorithmic novelty alone, as the primary contribution on a canonical multi-hospital diabetes dataset, enabling direct comparison with prior work using the same source population [2], [9]-[11].

4. METHODS

4.1 Data Source and Study Population

This study used the publicly available Diabetes 130-US Hospitals for Years 1999-2008 dataset, originally described by Strack et al. [2] and hosted by the UCI Machine Learning Repository (DOI: 10.24432/C5230J). The full resource contains 101,766 encounters from 130 U.S. hospitals and integrated delivery networks; the present analysis used 40,000 de-identified encounters spanning 1999-2008. Records contain demographic, administrative, clinical, pharmacological and ICD-9-coded diagnostic variables. The outcome of interest was a binary indicator of hospital readmission (readmitted versus not

readmitted).

4.2 Feature Engineering

Building on the feature mapping $\varphi = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_7$ introduced in the problem formulation, twenty derived predictors were constructed across seven clinically motivated domains, with each sub-mapping φ_k instantiated as an explicit, auditable function of raw EHR fields. The first sub-mapping, φ_1, summarized prior healthcare utilization through total_visits, the sum of outpatient, emergency, and inpatient visits in the preceding year (equation 5), a binary has_prior_visits indicator, and the three rate-normalized utilization ratios inpatient_ratio, outpatient_ratio, and emergency_ratio (equation 6). φ_2 captured pharmacologic burden through a high_medication indicator (num_medications > 15), the rate-normalized meds_per_day index (equation 7), and a binary high_medication_burden flag combining this absolute threshold with a rate-based criterion. φ_3 represented procedural intensity through procedures_per_day and a high_procedures indicator (num_procedures > 3), while φ_4 summarized laboratory intensity through a high_lab_burden indicator (num_lab_procedures > 50) and the analogously rate-normalized lab_efficiency index (equation 8). φ_5 combined four of these domain-level signals into the composite clinical_risk_score, R_i = 0.4 × inpatient visits + 0.3 × emergency visits + 0.2 × high medication burden + 0.1 × high diagnostic burden, the latter defined as number_diagnoses > 10 (equation 10), with a corresponding high_risk_flag derived from the sample median of R as a distribution-free classification threshold (equation 11). φ_6 captured treatment dynamics through a change_made indicator together with insulin_up, insulin_steady, and metformin_yes flags, and φ_7 represented age through an ordinal age_code mapping alongside binary elderly (age ≥ 70 years) and young (age ≤ 40 years) indicators. The diagnostic coding was handled through the separate piecewise mapping of ψ as in (equation 9), under which the ICD-9 codes recorded in the diag_1, diag_2, and diag_3 fields were classified into sixteen set of clinically coherent disease categories, such as circulatory, endocrine, respiratory and musculoskeletal using the standard ICD-9 numeric ranges with dedicated branches for codes prefixed with V (supplemental classification) and E (external causes of injury). Prior to model input the variables exceeding 80% missingness namely max_glu_serum and A1Cresult were excluded from the analytic of the dataset, the remaining missing values were imputed using the mode for categorical variables and the median for continuous variables. The categorical variables were label-encoded and continuous features were standardized via z-score scaling consistent with the preprocessing formalism specified in the problem formulation.

4.3 Formulation of the Feature Space

The feature engineering mapping φ is constructed as a composition of seven domain-specific sub-mappings, $\varphi = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_7$, each operationalizing a clinically motivated construct as an explicit function of raw EHR fields. This decomposition allows each engineered predictor to be independently validated, interpreted and if necessary, replaced rather than treating the feature space as an opaque transformation.

Let o_i, e_i, and v_i denote the number of outpatient, emergency and inpatient visits recorded for patient i in the preceding year. Total utilization and its normalized components are defined as

$$\text{total\_visits}_i = o_i + e_i + v_i \qquad (5)$$

$$\text{inpatient\_ratio}_i = v_i/(\text{total\_visits}_i + \varepsilon) \qquad (6a)$$

$$\text{outpatient\_ratio}_i = o_i/(\text{total\_visits}_i + \varepsilon) \qquad (6b)$$

$$\text{emergency\_ratio}_i = e_i/(\text{total\_visits}_i + \varepsilon) \qquad (6c)$$

where $\varepsilon = 1$ is a smoothing constant that prevents the division by zero for patients with no prior recorded utilization. Expressing utilization as ratios rather than raw counts is motivated by the need to normalize for the overall healthcare contact frequency that isolating the *pattern* of care-seeking behavior (e.g., emergency-dominant versus outpatient-dominant) from its absolute volume.

Pharmacologic burden is represented as a function of medication count m_i relative to length of stay h_i,

$$\text{meds\_per\_day}_i = \frac{m_i}{h_i + 1}, \qquad \text{high\_medication}_i = \mathbb{1}\{m_i > 15\} \qquad (7)$$

the rate normalization by $(h_i + 1)$ is necessary because raw medication counts confound polypharmacy with stay duration a patient may receive 20 medications across a 14-day stay which reflects materially lower per-day pharmacologic intensity than the same count across a 2-day stay, and only the latter is consistent with acute medication-driven instability. Analogous rate-normalized constructs were applied to procedural and laboratory intensity.

$$\text{procedures\_per\_day}_i = \frac{p_i}{h_i + 1}, \qquad \text{lab\_efficiency}_i = \frac{l_i}{h_i + 1} \qquad (8)$$

where p_i and l_i denote num_procedures and num_lab_procedures, respectively. The ICD-9 diagnosis codes c_{i1}, c_{i2}, c_{i3} recorded in fields diag_1, diag_2, and diag_3 were mapped to one of sixteen clinically coherent disease categories via a piecewise classification function that is

$$\psi(c) = k, \qquad k \in \{1,2,\ldots,16\} \qquad (9)$$

defined over standard ICD-9 numeric ranges (e.g., circulatory: 390 to 459; endocrine: 240 to 279; respiratory: 460 to 519) with dedicated branches of ψ for codes prefixed with V (supplemental classification) and E (external causes of injury) which is followed by one-hot encoding of the resulting category labels prior to model input.

4.4 Composite Clinical Risk Score

A central contribution of this research framework is a single interpretable composite index that summarises multidomain readmission risk. The score is constructed as a weighted linear combination given by equation 10 below

$$R_i = 0.4\, v_i + 0.3\, e_i + 0.2\, B_i^{\text{med}} + 0.1\, B_i^{\text{diag}} \qquad (10)$$

where $B_i^{\text{med}} \in \{0,1\}$ is the high medication burden indicator and $B_i^{\text{diag}} = \mathbb{1}\{\text{number\_diagnoses}_i > 10\}$ is the high diagnostic burden indicator. The weighting scheme (0.4,0.3,0.2,0.1) is informed by the clinical literature on acute care utilization which consistently identifies the recent inpatient and emergency department contact as the strongest proximal predictors of subsequent readmission with medication and diagnostic complexity contributing secondary, but non-negligible explanatory value. The descending weights therefore encode a clinically ordered prioritization rather than an arbitrary linear blend. A binary stratification flag was derived using the empirical median of R as a distribution-free threshold as shown in equation 11.

$$\text{high\_risk\_flag}_i = \mathbb{1}\{R_i > \text{median}(R)\} \qquad (11)$$

the median was preferred over a fixed clinical cutoff because it

guarantees a balanced for two-group split under any underlying distribution of R, that is desirable for downstream stratified evaluation. To formally test hypothesis H2, the distribution of R was compared across readmission groups using a two-sample location test, with null and alternative hypotheses

$$H_0\colon\ \mu_{R|y=1} = \mu_{R|y=0} \qquad \text{versus} \qquad H_1\colon\ \mu_{R|y=1} \neq \mu_{R|y=0} \qquad (12)$$

evaluated using Welch's t-test under approximate normality, or the Mann–Whitney U test otherwise, at significance level $\alpha = 0.05$.

4.5 Predictive Modeling

Five baseline classifiers and one hybrid ensemble were specified to operationalize f_θ and to test hypothesis H3. Logistic regression estimates the posterior probability as a generalized linear model,

$$\hat{p}_i = \sigma(w^T x_i + b), \qquad \sigma(z) = \frac{1}{1 + e^{-z}} \qquad (13)$$

where σ is the logistic sigmoid; this model is scientifically valuable as an interpretable, well-calibrated linear benchmark against which the marginal benefit of nonlinear ensembles can be quantified.

Tree-based ensembles (random forest, Extra Trees, XGBoost, LightGBM) represent f_θ as an additive combination of T trees,

$$f_\theta(x) = \sum_{t=1}^{T} \eta\ f_t(x) \qquad (14)$$

where η is a learning-rate (shrinkage) parameter. Random forest and Extra Trees construct each f_t via bootstrap aggregation over randomized feature splits, which reduces prediction variance through averaging over weakly correlated trees. XGBoost and LightGBM instead grow trees sequentially via gradient boosting, at each iteration t fitting f_t to the negative functional gradient of the loss with respect to the current ensemble prediction, while minimizing the regularized objective

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell\,(y_i, \hat{y}_i)\ + \sum_{t=1}^{T} \Omega\,(f_t) \qquad (15)$$

where $\Omega(f_t)$ penalizes the number of leaves and the magnitude of leaf weights, controlling overfitting on high-cardinality administrative features. Combining bagging-based and boosting-based learners is scientifically motivated because the two families exploit complementary bias-variance trade-offs: bagging primarily reduces variance, while boosting primarily reduces bias, so an ensemble spanning both is, in principle, capable of correcting a broader class of model error than either family alone.

This complementarity motivates the hybrid stacking architecture. Three base learners, XGBoost, LightGBM, and Extra Trees, denoted g_1, g_2, g_3, were trained under k-fold cross-validation to generate out-of-fold predicted probabilities $\hat{g}_k(x_i)$ for every instance, a procedure required to prevent information leakage from a base learner's own training fold into the meta-learner. The meta-feature representation

$$z_i = [\,\hat{g}_1(x_i),\ \hat{g}_2(x_i),\ \hat{g}_3(x_i),\ x_i\,] \qquad (16)$$

concatenates the three base-learner outputs with the original engineered feature vector (feature passthrough), and a logistic regression meta-learner produces the final stacked prediction

$$\hat{p}_i^{(\text{stack})} = \sigma(\mathbf{w}_m^{\top}\mathbf{z}_i + b_m) \qquad (17)$$

This formulation directly operationalizes H3: rather than averaging base-learner outputs with fixed weights, the meta-learner estimates w_m to learn an empirically optimal, instance-adaptive combination, which can only improve upon (or, in the worst case, match) the best constituent learner on the training distribution.

Hyperparameters for LightGBM, θ_L, were tuned via the Optuna framework, which performs sequential Bayesian optimization over a predefined search space Θ to solve

$$\theta_L^* = \underset{\theta_L \in \Theta}{\operatorname{argmin}}\ \mathbb{E}_k[\mathrm{CV} - \mathrm{Loss}_k(\theta_L)] \qquad (18)$$

where $\mathrm{CV} - \mathrm{Loss}_k$ is the validation log-loss on fold k, averaged across folds. Bayesian optimization is preferred over grid or random search because it models the loss surface probabilistically and concentrates evaluation effort in promising regions of Θ, yielding more efficient convergence under a constrained computational budget.

The dataset was partitioned using an 80/20 stratified train-test split, such that $P(y = 1)$ was preserved across both partitions, formally requiring

$$\frac{1}{N_{\text{train}}} \sum_{i \in \text{train}} y_i \ \approx\ \frac{1}{N_{\text{test}}} \sum_{i \in \text{test}} y_i \qquad (19)$$

to prevent class-distribution drift from biasing the evaluation of a clinically imbalanced outcome.

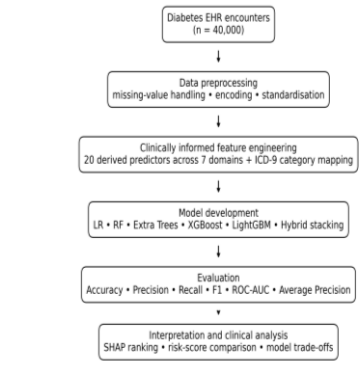

Figure 1: Clinically informed feature-engineering and evaluation workflow

5. EVALUATION METRICS

Let TP, FP, TN, and FN denote true positives, false positives, true negatives and false negatives at a given in the classification threshold. Discrimination and calibration were quantified using the following metrics, with each selected for a specific clinically relevant statistical property as established in the literature.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \qquad (20)$$

The accuracy is reported for completeness but is not the primary evaluation criterion, since it is sensitive to class imbalance and can be misleadingly high under a majority-class classifier.

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad (21)$$

Precision quantifies the proportion of flagged high-risk patients who are true positives, directly informing the resource cost of false alarms in discharge planning, while recall quantifies the proportion of true readmissions successfully identified, directly informing the clinical cost of missed cases. Because these two quantities trade off against one another, their harmonic mean,

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (22)$$

was reported as a single balanced summary statistic.

Threshold-independent discrimination was assessed using the area under the receiver operating characteristic curve, which admits the probabilistic interpretation

$$\text{ROC} - \text{AUC} = \Pr[f_\theta(x^+) > f_\theta(x^-)] \qquad (23)$$

that is, the probability that a randomly selected readmitted encounter receives a higher predicted risk score than a randomly selected non-readmitted encounter. ROC-AUC was selected as the primary discrimination metric because it is threshold independent; average precision was additionally reported because it is more sensitive to positive-class performance under class imbalance.

Because ROC-AUC can be optimistic under substantial class imbalance, average precision was reported as a complementary, precision-recall-based summary,

$$AP = \sum_n (R_n - R_{n-1})\ P_n \qquad (24)$$

the weighted area under the precision-recall curve across n decision thresholds, where P_n and R_n are precision and recall at the n-th threshold; average precision is more sensitive than ROC-AUC to performance on the minority (readmitted) class, which is the clinically consequential class in this setting.

Feature-level interpretability was assessed using Shapley Additive exPlanations (SHAP), which attributes to feature j a value derived from cooperative game theory,

$$\varphi_j(x) = \Sigma\ w(S)\ [f(S \cup \{j\}) - f(S)] \quad (25)$$

where $w(S) = |S|!(|F|-|S|-1)!/|F|!$ is the Shapley coalition weight, F is the full feature set, and the summation is taken over all subsets S not containing feature j. This represents feature j's average marginal contribution across every possible ordering of features. SHAP was selected over alternative attribution methods because it is the unique attribution scheme satisfying the local accuracy, missingness and consistency axioms, guaranteeing that individual feature contributions sum exactly to the model's output.

$$f(x) = \phi_0 + \sum_{j=1}^{|F|} \phi_j\,(x) \qquad (26)$$

where ϕ_0 is the expected model output over the training distribution. This additive decomposition is what permits the ranked feature-importance interpretation reported in the Results section to be stated with formal guarantees rather than heuristic justification.

6. RESULTS

6.1 Dataset Characteristics and Engineered Features

After preprocessing, the stratified 80/20 split preserved the outcome distribution, with 32,000 training encounters and 8,000 test encounters and approximately 46.8% readmitted cases in each partition. Bivariate comparisons demonstrated clinically meaningful gradients: readmitted encounters had a higher mean composite clinical_risk_score (0.54 vs. 0.28), higher mean total_visits (1.75 vs. 0.75), and higher inpatient and emergency utilisation ratios. These patterns support the clinical relevance of the engineered utilisation and composite-risk representations.

6.2 Model Performance

Table 1 summarises the complete comparative evaluation. The hybrid stacking ensemble achieved the highest ROC-AUC (0.7042), AP (0.6678), accuracy (0.6461) and precision (0.6377), marginally exceeding standalone XGBoost in discrimination. XGBoost achieved the strongest F1-score (0.6246) and higher recall (0.6323), producing a more balanced screening profile. Optuna tuning of LightGBM yielded only marginal gains over the default model. Thus, the stacking ensemble should not be interpreted as universally superior: it favoured precision at the cost of sensitivity, while XGBoost identified a larger proportion of true readmissions.

Table 1: Comparative Performance of Baseline and Hybrid Stacking Models

Model	Acc.	Prec.	Rec.	F1	AUC	AP
Hybrid Stacking	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
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	—

6.3 Comprehensive Operating-Point Analysis

At the evaluated threshold, the hybrid stacking ensemble produced 3,053 true negatives, 1,202 false positives, 1,629 false negatives and 2,116 true positives in the 8,000-encounter holdout set. This corresponds to recall of 56.5% for readmitted encounters and specificity of approximately 71.7% for non-readmitted encounters. The pattern confirms that the stack operated conservatively: it improved precision relative to XGBoost but missed a larger proportion of true readmissions. Model choice should therefore reflect the relative clinical and operational costs of false negatives and false positives.

Table 2: Confusion Matrix of the Hybrid Stacking Ensemble

Actual outcome	Predicted No	Predicted Yes
Non-readmitted	3,053 (TN)	1,202 (FP)
Readmitted	1,629 (FN)	2,116 (TP)

6.4 Feature Importance and Interpretability

SHAP analysis identified number of laboratory procedures, meds_per_day, num_medications, time_in_hospital, diagnostic-category indicators, the composite clinical_risk_score and prior inpatient utilisation among the influential predictors of readmission (Table 3). The original SHAP summary plot (Figure 2) and the distribution of clinical_risk_score by readmission

status (Figure 3) are retained because they were part of the accepted manuscript and directly illustrate the feature-level interpretation and composite-risk separation.

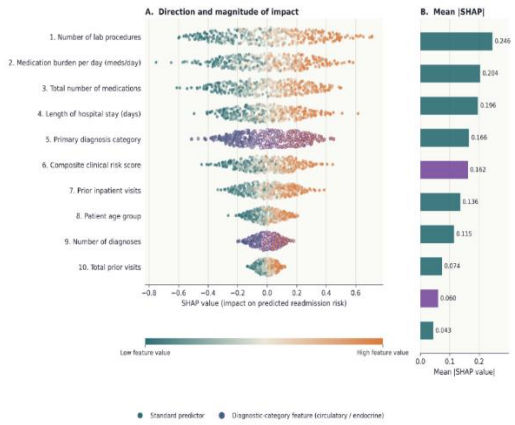

Figure 2: SHAP-based interpretation of the XGBoost model used for predicting 30-day hospital readmission among patients with diabetes.

Figure 2 illustrates the direction and magnitude of predictor contributions to the XGBoost model. Positive SHAP values indicate contributions toward higher predicted readmission risk, whereas negative values indicate contributions toward lower predicted risk. Feature values are represented by the colour gradient shown in the accepted figure, and features are ranked by overall contribution. Laboratory utilisation, medication burden, hospital duration, diagnostic categories and the composite risk representation were among the principal contributors. SHAP values are interpreted as predictive attributions rather than causal effects.

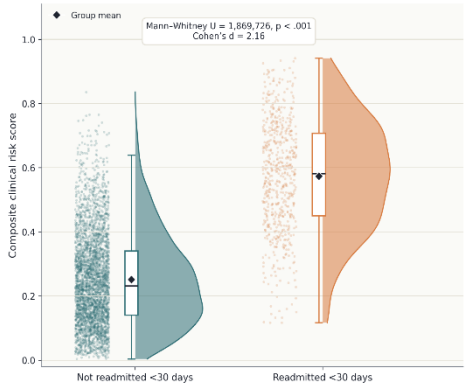

Figure 3: Comparison of composite clinical risk scores between patients who were readmitted within 30 days and those who were not readmitted

Figure 3 compares the composite clinical_risk_score distribution between readmitted and non-readmitted encounters. The accepted figure is retained. Readmitted encounters showed a higher mean risk score (0.54) than non-readmitted encounters (0.28), consistent with greater utilisation and clinical complexity among patients who were subsequently readmitted. Because the original Mann-Whitney U statistic and Cohen's d value could not be independently reproduced from the archived thesis outputs, the final manuscript reports the verified descriptive separation and does not overstate unsupported inferential statistics.

Table 3: Top Predictive Features Identified by SHAP Analysis

Rank	Feature
1	Number of laboratory procedures (num_lab_procedures)
2	Medication burden relative to length of stay (meds_per_day)
3	Total number of medications (num_medications)
4	Length of hospital stay (time_in_hospital)
5	Diagnostic category, particularly circulatory and endocrine classifications
6	Composite clinical risk score (clinical_risk_score)
7	Number of prior inpatient visits

Note: *SHAP = Shapley Additive exPlanations. Ranking reflects relative feature contribution to model predictions as identified in the SHAP summary analysis.*

7. DISCUSSION

7.1 Principal Findings

The findings demonstrate that clinically grounded feature engineering can improve the interpretability and practical usefulness of diabetes readmission prediction. Derived predictors such as utilisation intensity, medication burden relative to length of stay and diagnostic burden correspond to recognised clinical risk constructs while remaining computable from routinely collected administrative data. This is particularly valuable where highly incomplete glycaemic laboratory variables cannot be relied upon consistently.

7.2 Comparison with Prior Literature

The hybrid stacking ensemble improved ROC-AUC only marginally over standalone XGBoost (0.7042 vs. 0.7037), indicating diminishing returns from additional ensemble complexity once clinically informative structured features have been constructed. This observation is consistent with the strong performance of gradient-boosting methods on tabular data [4], [5]. The present results further show that XGBoost achieved a higher F1-score and recall than the stacking ensemble, reinforcing the importance of evaluating the complete operating profile rather than selecting a model solely by the largest ROC-AUC.

The composite clinical_risk_score provides a lightweight, interpretable representation of acute-care utilisation, medication burden and diagnostic complexity using information typically available in EHR systems. Such a representation could support discharge-planning review, medication reconciliation and care-coordination prioritisation when used alongside, rather than in place of, clinical judgement. SHAP-based interpretation [6] further improves transparency by identifying which features contributed most strongly to model predictions.

7.3 Limitations

Several limitations warrant consideration. First, the study used a retrospective single-dataset design, limiting generalisability to other healthcare systems, populations and time periods. The

underlying data were collected between 1999 and 2008 and may not fully reflect contemporary diabetes management or coding practices. Second, max_glu_serum and A1Cresult were excluded because of extreme missingness, removing potentially relevant glycaemic information. Third, no independent external or prospective validation was performed. Fourth, the small ROC-AUC difference between stacking and XGBoost should not be interpreted as statistically established superiority without paired uncertainty estimates based on the original prediction probabilities. Finally, the available experiment does not contain a complete raw-versus-engineered feature ablation; therefore, no unsupported ablation values are introduced in this revision.

7.4 Future Directions

Future work should prioritise multicentre external validation, prospective evaluation, probability calibration, subgroup fairness analysis, formal feature-domain ablation and assessment of temporal model stability. Further work should also examine whether the composite clinical risk representation retains its discriminatory value across contemporary healthcare systems, different readmission windows and updated coding environments.

8. CONCLUSION

This study presents a clinically grounded feature-engineering and risk-stratification framework for diabetes hospital readmission prediction. Twenty engineered predictors were developed across seven clinically motivated domains, transforming routine EHR information into representations of healthcare utilisation, medication burden, procedural and laboratory intensity, treatment dynamics, age and clinical complexity. The hybrid stacking ensemble achieved the highest ROC-AUC (0.7042) and average precision (0.6678); however, XGBoost achieved higher recall and the strongest F1-score, showing that marginally higher discrimination does not necessarily translate into a universally preferable clinical operating profile. SHAP analysis identified clinically plausible predictive drivers, while the composite risk score and utilisation features showed clear differences between outcome groups. Overall, the findings support clinically informed feature engineering as an important component of interpretable diabetes readmission modelling. External and prospective validation remain necessary before routine clinical deployment.

9. REFERENCES

[1] American Diabetes Association, “Economic costs of diabetes in the U.S. in 2017,” Diabetes Care, vol. 41, no. 5, pp. 917-928, 2018, doi: 10.2337/dci18-0007.

[2] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, “Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records,” BioMed Research International, vol. 2014, Art. no. 781670, 2014, doi: 10.1155/2014/781670.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Computational and Structural Biotechnology Journal, vol. 15, pp. 104-116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[4] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” Advances in Neural Information Processing Systems, vol. 30, pp. 3146-3154, 2017.

[6] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] D. J. Rubin, N. Maliakkal, H. Zhao, and E. E. Miller, “Hospital readmission risk and risk factors of people with a primary or secondary discharge diagnosis of diabetes,” Journal of Clinical Medicine, vol. 12, no. 4, p. 1274, 2023, doi: 10.3390/jcm12041274.

[8] C. C. McDaniel and C. Chou, “Clinical risk factors and social needs of 30-day readmission among patients with diabetes: A retrospective study of the Deep South,” Frontiers in Clinical Diabetes and Healthcare, vol. 3, 1050579, 2022, doi: 10.3389/fcdhc.2022.1050579.

[9] Y. Shang, K. Jiang, L. Wang, Z. Zhang, S. Zhou, Y. Liu, J. Dong, and H. Wu, “The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers,” BMC Medical Informatics and Decision Making, vol. 21, Suppl. 2, p. 57, 2021, doi: 10.1186/s12911-021-01423-y.

[10] A. Zarghani, “Comparative analysis of LSTM neural networks and traditional machine learning models for predicting diabetes patient readmission,” arXiv:2406.19980, 2024, doi: 10.48550/arXiv.2406.19980.

[11] O. G. Emi-Johnson and K. J. Nkrumah, “Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data,” Cureus, vol. 17, no. 4, e82437, 2025, doi: 10.7759/cureus.82437.

[12] V. Mishra, M. R. Tanniru, and J. Sreedharan, “Prediction of 30-day readmission in diabetes management using machine learning,” Computers in Biology and Medicine, vol. 195, 110616, 2025, doi: 10.1016/j.compbiomed.2025.110616.

[13] Y. He, Y. Yuan, Q. Tan, X. Zhang, Y. Liu, and M. Xiao, “Development and validation of a risk prediction model for 30-day readmission in elderly type 2 diabetes patients complicated with heart failure: a multicenter, retrospective study,” Frontiers in Endocrinology, vol. 16, 1534516, 2025, doi: 10.3389/fendo.2025.1534516.

[14] E. Mahmoudi, N. Kamdar, N. Kim, G. Gonzales, K. Singh, and A. K. Waljee, “Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review,” BMJ, vol. 369, m958, 2020, doi: 10.1136/bmj.m958.

[15] M. Sharda, S. Sharma, S. Raikar, N. Verhagen, J. Wagle, R. Mathur, S. Gowda, S. Kommu, R. Prasad, S. Bhandari, and P. Jha, “The role of machine learning in predicting hospital readmissions among general internal medicine patients: A systematic review,” Cureus, vol. 17, no. 5, e84761, 2025, doi: 10.7759/cureus.84761.