iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84701

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

A Reliability-Aware and Interpretable Machine Learning Framework for Diabetes Prediction Using Structured Clinical Data

Ravikumar V[1], Dr. S. Sasirekha[2]
[1]*Post-Graduate Student, NITTTR Chennai, India*
[2]*Associate Professor, NITTTR Chennai, India*

Abstract: ***Diabetes prediction plays an important role in re-ducing long-term health risks by enabling early medical interven-tion. Although machine learning models have been widely applied to this task, many existing studies emphasise predictive accuracy while giving comparatively little attention to the reliability, interpretability, and stability of the resulting decisions. This paper develops a reliability-aware and interpretable machine learning framework for diabetes prediction from structured clinical data. Three complementary models—Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost)—are trained on the Pima Indians Diabetes dataset so that both simple linear and complex non-linear relationships are captured. Beyond conventional discrimination metrics, the reliability of the predicted probabilities is quantified using the Brier score and reliability (calibration) diagrams. Interpretability is addressed with SHapley Additive exPlanations (SHAP) at both the global (cohort) and local (individual patient) levels. Because different models frequently emphasise different predictors, we formalise a Feature Consistency Index (FCI) that quantifies the cross-model agreement of SHAP-derived feature importance and combines it with normalised importance into a single ranking score. Finally, a perturbation-based robustness analysis measures the sensitivity of each model's output to small changes in the input record. Experi-mentally, XGBoost achieves the highest discrimination (accuracy*** *0.7597*, ***ROC-AUC*** *0.8374*), ***whereas Random Forest attains the best-calibrated probabilities (Brier score*** *0.1646*), ***demonstrating that discrimination and reliability are not interchangeable. The FCI identifies Glucose and BMI as simultaneously the most influential and the most consistently attributed predictors, while Blood Pressure and Skin Thickness are both weak and unstable. Under a*** *5%* ***Gaussian perturbation of a representative patient record, the linear and bagged models shift by less than*** *0.01* ***in predicted probability, whereas the boosted model shifts by*** *0.0386*, ***revealing an accuracy–stability trade-off that a purely accuracy-driven evaluation would not expose.***

Index Terms: ***Diabetes prediction, explainable artificial intel-ligence, SHAP, model calibration, Brier score, feature stability, robustness analysis, clinical decision support.***

I. INTRODUCTION

Diabetes mellitus is among the most prevalent chronic metabolic disorders worldwide, and its prevalence con-tinues to rise across both high- and low-income regions [1]. Its clinical importance lies less in the immediate symptoms of sustained hyperglycaemia than in the long-term microvascu-lar and macrovascular complications—retinopathy, nephropa-thy, neuropathy, and cardiovascular disease—that accumulate

Ravikumar V and Dr. S. Sasirekha are with the Department of Computer Science and Engineering, National Insitute of Technical Teachers Training and Research, Chennai, 600 113, India (e-mail: ravi.sitthaandi@gmail.com; sasirekha@nitttrc.edu.in). silently over years. Clinical evidence indicates that identifying at-risk individuals *before* overt disease onset, and intervening through lifestyle modification or pharmacotherapy, materially reduces the incidence of these complications, so computational screening tools that flag high-risk individuals from routinely collected measurements are of considerable practical value.

Machine learning (ML) has been applied extensively to this problem [2]–[4]. Reported studies typically train one or more classifiers on structured tabular data and report discrimination metrics such as accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (ROC-AUC), with progress measured as an increase in accuracy achieved by a more expressive model, a more aggressive resampling scheme, or a more thorough hyper-parameter search.

We argue that this template, while necessary, is insufficient for a clinical screening context, for three distinct reasons.

First, discrimination is not reliability. A model may rank patients correctly while producing probability estimates that are systematically over- or under-confident. In a screening workflow the predicted probability is not an intermediate quantity: it is the object on which a referral threshold is set and against which a clinician calibrates their own judgement. A model reporting 0.9 for a cohort in which only 60% of members are diabetic is misleading even if its ranking, and hence its AUC, is excellent. Calibration is therefore a first-class requirement, not a diagnostic afterthought [6], [7], [10]. *Second, an unexplained prediction is difficult to act upon.*

A risk score without a rationale gives the clinician no basis on which to accept, question, or override it, and no indication of which modifiable factor drives the risk. Post-hoc attribution methods, most prominently SHapley Additive exPlanations (SHAP) [14], [15], decompose an individual prediction into additive per-feature contributions grounded in cooperative game theory [16], [17].

Third—the gap this work targets—explanations are them-selves model-dependent. Models of comparable accuracy trained on the same data frequently disagree about *why* a patient is at risk. If Logistic Regression attributes risk princi-pally to Glucose and Pregnancies while a boosted ensemble attributes it to Glucose, BMI, and Age, the explanation the clinician receives is an artefact of an arbitrary modelling choice. Work on the stability of feature selection [25], [26] and the fragility of interpretation methods [23], [24] establishes that this instability is real and consequential, but it is rarely quantified in applied clinical ML studies, which typically report the importances of a single chosen model.

A. *Contributions*

This paper develops and evaluates an integrated framework that treats accuracy, reliability, interpretability, and stability as four coordinate axes of model assessment rather than as a primary objective plus three optional diagnostics. The specific contributions are as follows.

1) We construct a reproducible diabetes-prediction pipeline on the Pima Indians Diabetes dataset that com-bines physiologically-motivated missing-value treat-ment, stratified evaluation, and three complementary classifiers spanning the linear–bagged–boosted spec-trum.

2) We evaluate probabilistic reliability explicitly using the Brier score and reliability diagrams, and show empiri-cally that the model with the best discrimination is not the model with the best-calibrated probabilities on this dataset.

3) We formalise a Feature Consistency Index (FCI), a coefficient-of-variation-based measure of cross-model agreement in SHAP attribution, and combine it with normalised importance into a composite ranking that rewards predictors which are both influential and con-sistently identified.

4) We apply SHAP at the individual-patient level and demonstrate that the cohort-level ranking does not trans-fer to a specific patient, motivating per-patient rather than global explanation in a screening interface.

5) We conduct a perturbation-based robustness analysis that exposes an accuracy–stability trade-off invisible to conventional metrics: the most accurate model is also the most sensitive to small input changes.

The remainder of this paper is organised as follows. Sec-tion II reviews related work. Section III describes the dataset, preprocessing, models, and the four evaluation axes, including the formal definition of the FCI. Section IV presents the experimental results. Section V interprets these results and Section VI states the limitations of the study. Section VII concludes.

II. RELATED WORK

A. *Machine Learning for Diabetes Prediction*

The Pima Indians Diabetes dataset, introduced by Smith *et al.* [5] in connection with the ADAP learning algorithm, has become the canonical benchmark for tabular diabetes prediction. Its enduring use is a product of its accessibility and modest size; its limitations—a single, demographically narrow cohort, only eight predictors, and substantial physiologically implausible missingness—are equally well known and are revisited in Section VI.

Broad surveys of ML in diabetes research [2] document the application of essentially the full classifier catalogue to this problem. Zou *et al.* [3] compared decision trees, random forests, and neural networks, additionally examining the effect of dimensionality reduction; Maniruzzaman *et al.* [4] evaluated a range of classifiers combined with several feature-selection strategies and reported the sensitivity of the final result to that combination—an early indication that conclusions drawn from a single pipeline are fragile. Reported accuracies on this dataset vary widely, and much of that variance is attributable to methodological choices that are not always reported in sufficient detail: whether implausible zeros were treated as missing, whether resampling preceded the train–test split, whether the split was stratified, and whether hyper-parameters were tuned on the test partition. Where preprocessing is handled rigorously and evaluation is genuinely held out, accuracies of 0.75–0.78 and ROC-AUC values of 0.82–0.85 are typical, and the results of Section IV sit within this range.

B. Probability Calibration in Clinical Prediction

The distinction between discrimination and calibration is long established in the clinical prediction literature [7]. Van Calster *et al.* [6] describe calibration as "the Achilles heel of predictive analytics," noting that miscalibrated models can cause clinical harm even when their discrimination is excel-lent, because the decision thresholds applied to their outputs are then misaligned with true risk. The TRIPOD reporting guideline [8] accordingly requires that both discrimination and calibration be reported for any clinical prediction model.

The Brier score [9] is the standard proper scoring rule for this purpose; being proper, it is minimised only by the true probability distribution and therefore jointly rewards dis-crimination and calibration, with reliability diagrams provid-ing the corresponding graphical diagnostic. Niculescu-Mizil and Caruana [10] showed that model families exhibit char-acteristic calibration distortions—maximum-margin methods pushing probabilities toward the extremes, bagged ensembles compressing them toward the centre. Post-hoc recalibration by Platt scaling [11] or isotonic regression [12] can mitigate these distortions, and the deep learning literature has revisited the same problem for over-confident networks [13]. Despite this well-developed base, calibration remains under-reported in applied ML studies of diabetes prediction; the present work treats it as a primary axis.

C. Interpretability and the Stability of Explanations

Post-hoc explanation methods aim to render an opaque model's behaviour intelligible without altering the model. LIME [18] fits a sparse local surrogate around the instance of interest; SHAP [14] instead assigns each feature the Shapley value [16] of a cooperative game whose payoff is the model output, yielding the unique attribution satisfying local accu-racy, missingness, and consistency. TreeSHAP [15] makes this computation tractable in polynomial time for tree ensembles, which is what enables its routine use with Random Forest and XGBoost.

SHAP is not without well-documented caveats: its standard estimators assume feature independence, which is violated in correlated clinical data [20]; multiple non-equivalent for-mulations of "the" Shapley value exist [21]; and post-hoc explanations can be manipulated by an adversarial model [22]. Rudin [19] argues on these grounds that intrinsically inter-pretable models should be preferred for high-stakes decisions. A further concern, and the one most directly relevant here, is stability. Ghorbani *et al.* [23] showed that interpretations can be altered substantially by imperceptible input perturbations that leave the prediction unchanged, and Alvarez-Melis and Jaakkola [24] formalised robustness criteria for interpretability methods. In the adjacent feature-selection literature, Kalousis *et al.* [25] and Nogueira *et al.* [26] developed formal stability measures for subsets returned under data resampling.

The measure proposed in this paper is related to but distinct from these. Whereas the stability literature typically asks how a *single* algorithm's output varies under perturbation of the *data*, we ask how attribution for a *fixed* dataset varies across *different model families*. This is the practically relevant question when a study must choose one model to deploy and to explain: if the explanation is an artefact of that choice, its clinical value is limited. Molnar [27] discusses this cross-model disagreement qualitatively; the FCI defined in Section III-H makes it a quantity that can be computed, ranked, and reported.

III. MATERIALS AND METHODS

The framework proceeds in five stages. Raw clinical records pass through physiologically-motivated missing-value treat-ment and stratified partitioning; three model families are then trained and evaluated along four axes—discrimination, relia-bility, interpretability, and robustness—with the FCI acting as the bridge between the per-model explanations and a single consolidated feature ranking.

A. Dataset

All experiments use the Pima Indians Diabetes dataset [5], comprising 768 records of female patients of Pima Indian heritage aged at least 21 years. Each record contains eight predictors—number of pregnancies, plasma glucose concen-tration at 2 h in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin-fold thickness (mm), 2-h serum insulin (μU/mL), body mass index (kg/m^2), diabetes pedigree function, and age (years)—together with a binary outcome indicating a diagnosis of diabetes.

The cohort contains 500 negative and 268 positive cases, an imbalance ratio of approximately 1.87:1 (Fig. 1(a)). This imbalance is moderate but sufficient to make raw accuracy misleading: a degenerate classifier predicting "non-diabetic" for every patient attains 65.1% accuracy while providing no clinical value. All reported accuracies must therefore be read against this floor, and class-conditional metrics are reported throughout.

B. Missing Value Treatment

A well-documented characteristic of this dataset is that missing measurements are encoded as zero rather than as an explicit missing marker. For five of the eight predictors a value of zero is not merely unusual but physiologically impossible in a living patient: a plasma glucose concentration, diastolic blood pressure, triceps skin-fold thickness, serum insulin level, or body mass index of exactly zero cannot occur. Treating these encoded zeros as genuine measurements biases every downstream statistic—it depresses the estimated mean, inflates the estimated variance, and distorts the correlation structure that feature-importance analysis subsequently interprets. Accordingly, zero values in {Glucose, BloodPressure, SkinThickness, Insulin, BMI} were reinterpreted as missing. Fig. 1(b) shows the resulting missingness rates, which vary by two orders of magnitude across features: Glucose is affected in only 0.7% of records, whereas Insulin is affected in 48.7% and Skin Thickness in 29.6%. Missing values were then imputed with the median of the corresponding feature, the median being preferred to the mean because several of these distributions are strongly right-skewed. The choice to impute rather than to discard is deliberate: listwise deletion of records with any implausible zero would remove roughly half the cohort and would do so non-randomly, since missingness in Insulin and Skin Thickness is unlikely to be independent of patient characteristics. The implications of this choice, particularly for the two heavily affected features, are examined in Section VI.

C. Experimental Protocol

The cohort was partitioned into a training set of 614 records (80%) and a held-out test set of 154 records (20%) using stratified sampling, which preserves the outcome distribution in both partitions (34.85% positive in training, 35.06% positive in test). A fixed random seed was used throughout so that the partition, the model fits, and all reported figures are reproducible.

Features were standardised to zero mean and unit variance for Logistic Regression, which is sensitive to predictor scale through its regularisation penalty. The scaler was fitted exclusively on the training partition and then applied to the test partition, so that no information from the held-out data influences the transformation. The two tree ensembles are invariant to monotone rescaling of individual features and were therefore trained on the unscaled values.

Two properties of this protocol bear on the interpretation of the results. First, imputation medians were computed over the full cohort prior to partitioning, which introduces a mild optimistic bias; the effect is small for a univariate median but is not zero, and a strictly nested implementation would fit the imputer inside the training fold. Second, no hyper-parameter search was performed against the test partition—the values in Section III-D are fixed a priori—so the reported metrics are not inflated by selection over the evaluation data.

D. Predictive Models

Three model families were selected to span the space of inductive biases relevant to structured clinical data.

1) Logistic Regression: A regularised linear model in the log-odds of the outcome,

$$\log \frac{p(y=1 \mid \mathbf{x})}{1-p(y=1 \mid \mathbf{x})} = o_0 + \sum_{j=1}^{d} \beta_j x_j \qquad (1)$$

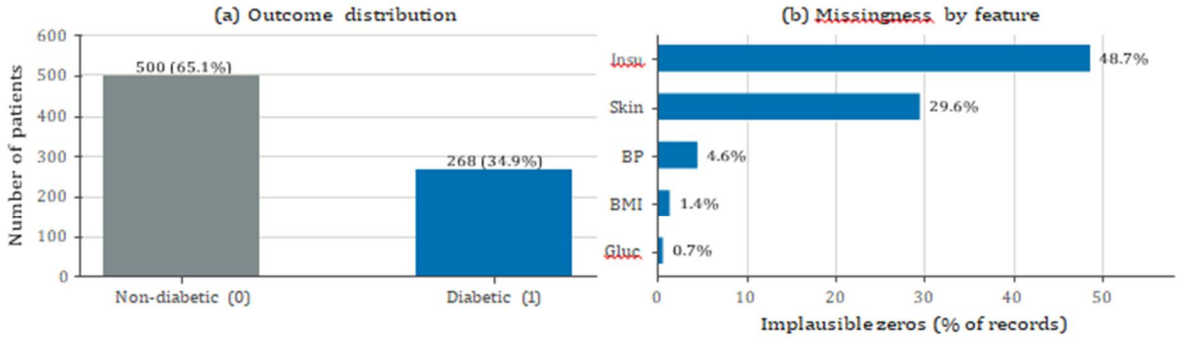


Fig. 1. (a) Outcome distribution: the 65.1% majority-class share is the accuracy floor against which all reported accuracies must be interpreted. (b) Proportion of records containing physiologically implausible zero values. Insulin and Skin Thickness are affected in 48.7% and 29.6% of records respectively, so their imputed values carry substantially less information than the remaining predictors.

fitted by penalised maximum likelihood with the default ℓ_2 penalty and a maximum of 1000 iterations. Logistic Regres-sion serves two roles here: it is the clinically conventional baseline, and because it is additive in the log-odds it provides an intrinsically interpretable reference against which the post-hoc explanations of the ensembles can be compared.

2) *Random Forest:* An ensemble of 200 decision trees [28] grown to a maximum depth of 6, each on a bootstrap resample of the training data and with feature subsampling at each split. The predicted probability is the average of the per-tree class frequencies. Class weights were set inversely proportional to class frequency to counteract the outcome imbalance. Aver-aging over decorrelated trees reduces variance and captures interactions the linear model cannot, at the cost of direct interpretability.
3) *XGBoost:* A gradient-boosted tree ensemble [29] of 300 estimators with maximum depth 4, learning rate 0.05, row subsampling of 0.8, and column subsampling of 0.8. Boosting fits each successive tree to the residual error of the current ensemble, typically yielding the strongest discrimination on tabular data but also the greatest capacity to fit local structure. Imbalance was handled through scale_pos_weight, set to the ratio of negative to positive training instances (400/214 ≈ 1.87), and log-loss was the training objective. The shallow depth, moderate learning rate, and subsampling ratios together act as regularisation appropriate to a dataset of fewer than a thousand records.

E. Axis 1: Discrimination

Discrimination was assessed with accuracy, per-class preci-sion, recall, and F_1 score, and the area under the ROC curve. Because the outcome is imbalanced and the costs of the two error types are asymmetric—a missed diabetic patient forgoes an intervention opportunity, whereas a false positive incurs unnecessary confirmatory testing—recall on the positive class is reported separately rather than absorbed into an aggregate score.

F. Axis 2: Reliability

Let $\hat{p}_i \in [0, 1]$ be the predicted probability of diabetes for test record i and $y_i \in \{0, 1\}$ the observed outcome. The Brier score [9] over N test records is

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2. \qquad (2)$$

Lower values are better. As a strictly proper scoring rule, BS is minimised in expectation only when $\hat{p}_i$ equals the true conditional probability, so it penalises both poor ranking and poor probability magnitude. It admits the classical decompo-sition into reliability, resolution, and uncertainty components, of which the reliability term captures calibration in the narrow sense.

As a reference point, a model predicting the training preva-lence $\bar{y}$ for every patient attains $\mathrm{BS} = \bar{y}(1-\bar{y}) \approx 0.227$ on a cohort of this prevalence, while a perfect model attains 0.

Calibration was additionally assessed with reliability di-agrams: test records were grouped into seven equal-width bins by predicted probability, and within each bin the mean predicted probability was plotted against the observed event frequency. A perfectly calibrated model traces the diagonal; excursions above it indicate under-confidence and below it over-confidence.

G. Axis 3: Interpretability

Feature attribution was computed with SHAP [14]. For a model f and instance $\mathbf{x}$, SHAP assigns to each feature j the value

$$\phi_j(f, \mathbf{x}) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f_{S\cup\{j\}}(\mathbf{x}) - f_S(\mathbf{x})\right], \qquad (3)$$

where F is the full feature set and f_S denotes the model restricted to subset S. Equation (3) is the Shapley value [16] of the cooperative game in which features are players and the payoff is the model output, and it is the unique attribution satisfying local accuracy, missingness, and consistency. Attri-butions are additive: $f(\mathbf{x}) = \phi_0 + \sum_j \phi_j(f, \mathbf{x})$, where ϕ_0 is the expected model output over the background distribution. The exact combinatorial sum in (3) is intractable for $|F|$ of realistic size, so model-specific estimators were used: the exact linear estimator for Logistic Regression, and TreeSHAP [15] for the two tree ensembles.

Global importance for model *m* and feature j was sum-marised as the mean absolute attribution over the *N* test records,

$$\bar{\phi}_j^{(m)} = \frac{1}{N} \sum_{i=1}^{N} \phi_j\left(f^{(m)}, \mathbf{x}_i\right). \qquad (4)$$

Local explanations for a single patient use the signed values $\phi_j(f^{(m)}, \mathbf{x})$ directly, which distinguishes factors that raise a patient's risk from those that lower it.

H. Axis 4: The Feature Consistency Index

Equation (4) yields one importance profile per model. The FCI consolidates the set $\mathbf{M} = \{\text{LR, RF, XGB}\}$ of these profiles into a per-feature agreement score.

For feature j, let μ_j and σ_j be the mean and (population) standard deviation of its importance across models,

$$\mu_j = \frac{1}{|\mathbf{M}|} \sum_{m \in \mathbf{M}} \bar{\phi}_j^{(m)}, \qquad \sigma_j = \sqrt{\frac{1}{|\mathbf{M}|} \sum_{m \in \mathbf{M}} \left(\bar{\phi}_j^{(m)} - \mu_j\right)^2}. \qquad (5)$$

The dispersion of the attributions relative to their magnitude is the coefficient of variation,

$$\mathrm{CV}_j = \frac{\sigma_j}{\mu_j + \varepsilon}, \qquad (6)$$

with $\varepsilon = 10^{-6}$ guarding against division by zero for a feature that every model ignores. The Feature Consistency Index is then the bounded transform

$$\mathrm{FCI}_j = \frac{1}{1 + \mathrm{CV}_j} \in (0, 1]. \qquad (7)$$

$\mathrm{FCI}_j = 1$ denotes perfect agreement ($\sigma_j = 0$); values approaching 0 denote attributions that vary widely relative to their mean. The transform is monotonically decreasing in CV_j and maps the unbounded coefficient of variation onto a bounded, directly comparable scale.

Consistency alone is not sufficient for ranking, since a feature that all three models agree is irrelevant would score highly. The FCI is therefore combined with importance. Mean importance is first min–max normalised across the *d* features,

$$\tilde{I}_j = \frac{\mu_j - \min_k \mu_k}{\max_k \mu_k - \min_k \mu_k} \in [0, 1], \qquad (8)$$

and the composite score is the convex combination

$$S_j = \alpha \tilde{I}_j + (1 - \alpha)\, \mathrm{FCI}_j, \qquad \alpha = 0.6. \qquad (9)$$

The weighting $\alpha = 0.6$ places the majority of the weight on importance while allowing consistency to act as a substantial tie-breaker; the sensitivity of the resulting ranking to this choice is examined in Section V. Features are ranked by S_j in descending order.

The same construction applies unchanged at the level of a single patient by replacing the cohort-averaged $\bar{\phi}_j^{(m)}$ in (4) with the absolute local attribution $|\phi_j(f^{(m)}, \mathbf{x})|$ for that pa-tient, yielding a per-patient consistency ranking.

TABLE I
DESCRIPTIVE STATISTICS AFTER MISSING-VALUE TREATMENT
(n = 768)

Feature	Mean	SD	Min	Median	Max
Pregnancies	3.85	3.37	0.00	3.00	17.00
Glucose	121.66	30.44	44.00	117.00	199.00
Blood Pressure	72.39	12.10	24.00	72.00	122.00
Skin Thickness	29.11	8.79	7.00	29.00	99.00
Insulin	140.67	86.38	14.00	125.00	846.00
BMI	32.46	6.88	18.20	32.30	67.10
Diabetes Pedigree Func.	0.47	0.33	0.08	0.37	2.42
Age	33.24	11.76	21.00	29.00	81.00

I. Axis 5: Robustness to Input Perturbation

Clinical measurements carry inherent variability arising from instrument precision, sample handling, biological fluc-tuation, and time of measurement; a prediction that changes materially under variation of this magnitude is not clinically dependable regardless of its aggregate accuracy. Robustness was therefore assessed by perturbing a representative patient record with multiplicative Gaussian noise. For each feature j of record $\mathbf{x}$, the perturbed value is

$$\tilde{x}_j = x_j + \eta_j, \qquad \eta_j \sim \mathrm{N}\left(0, (\lambda x_j)^2\right), \qquad (10)$$

with noise level $\lambda = 0.05$, corresponding to a 5% relative standard deviation. Scaling the noise by the feature value keeps the perturbation commensurate across predictors whose natural units differ by orders of magnitude. The sensitivity of model m is reported as the absolute change in predicted probability,

$$\Delta^{(m)} = f^{(m)}(\mathbf{x}) - f^{(m)}(\tilde{\mathbf{x}}). \qquad (11)$$

IV. RESULTS

A. Cohort Characteristics

Table I reports the descriptive statistics of the eight pre-dictors after missing-value treatment, and Fig. 2 shows the pairwise Pearson correlation structure.

Glucose exhibits the strongest marginal association with the outcome (r = 0.49), consistent with its diagnostic role—the outcome label is itself derived from glucose-based criteria, so this association is definitional as much as empirical—followed by BMI (r = 0.31) and Age (r = 0.24).

Among the predictors, the strongest correlations are Age–Pregnancies (r = 0.54) and BMI–Skin Thickness (r = 0.54), the latter reflecting that both quantify adiposity.

These correlations matter for the attribution results, because SHAP's standard estimators assume feature independence and can distribute credit arbitrarily between correlated predictors [20].

Although Glucose and BMI separate the outcome groups clearly in central tendency, their interquartile ranges overlap substantially between groups. This overlap is the fundamental reason no model here exceeds approximately 0.76 accuracy: the eight available predictors do not separate the classes cleanly, and no amount of model capacity can manufacture information the features do not contain.

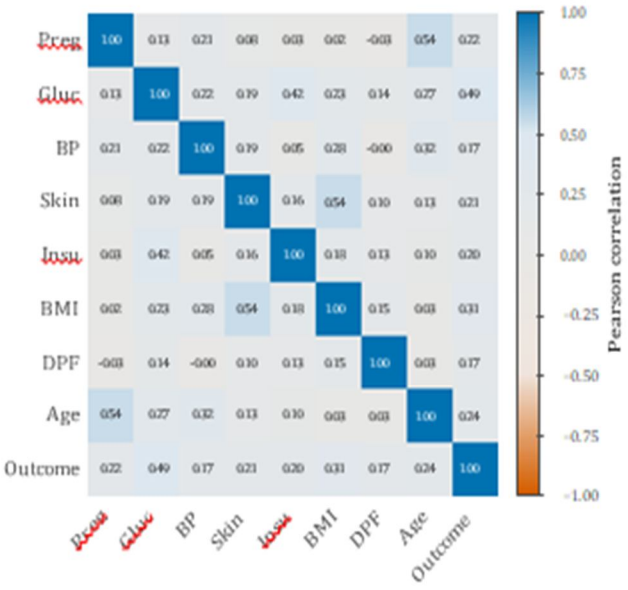


Fig. 2. Pearson correlation matrix after missing-value treatment. Glucose shows the strongest association with the outcome; the Age–Pregnancies and BMI–Skin Thickness pairs are the most strongly correlated predictors, which is relevant to the attribution analysis of Section IV-D.

B. Discriminative Performance

Table II reports held-out performance for the three models. XGBoost achieves the highest accuracy (0.7597) and the high-est ROC-AUC (0.8374), followed by Random Forest (0.7468, 0.8239) and Logistic Regression (0.7078, 0.8130). All three exceed the 65.1% majority-class floor, and the ordering is consistent with the expected capacity hierarchy of the three model families.

The aggregate figures, however, conceal a clinically im-portant difference in error profile. Random Forest attains the highest recall on the diabetic class (0.76), correctly identifying 41 of the 54 positive test cases, whereas Logistic Regression recovers only 27 of 54 (recall = 0.50)—it misses half of the patients the system exists to find. XGBoost sits between the two at 0.70 recall with higher precision (0.64 versus 0.61). This divergence is a direct consequence of the class imbalance: Logistic Regression achieves respectable accuracy largely by classifying the majority class well (0.82 recall on non-diabetic patients), which is of limited value in screening. If the objective is to minimise missed cases, the accuracy ordering of Table II inverts between the top two models.

C. Probabilistic Reliability

The final column of Table II reports Brier scores, and Fig. 3 shows the corresponding reliability diagrams. All three models improve substantially on the 0.227 no-skill reference, confirm-ing that the predicted probabilities carry real information.

The central finding of this subsection is that the reliability ordering does not match the discrimination ordering. Ran-dom Forest yields the lowest—that is, the best—Brier score (0.1646), narrowly ahead of XGBoost (0.1656), even though XGBoost is clearly superior on both accuracy and ROC-AUC. Logistic Regression is worst on this measure (0.1754) despite producing probabilities from a model whose functional form is expressly probabilistic.

This dissociation follows from the definitions. ROC-AUC depends only on the *ordering* of predicted probabilities and is invariant to monotone transformations of them, whereas the

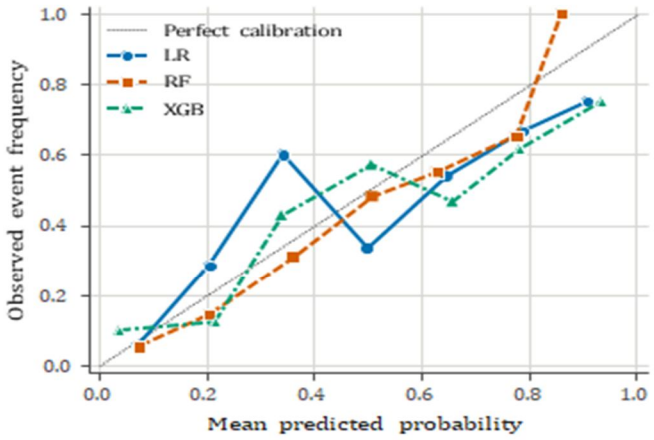


Fig. 3. Reliability diagrams on the held-out test set (seven equal-width bins). The dotted diagonal is perfect calibration; points above it indicate under-confidence and points below over-confidence. With 154 test records distributed over seven bins, individual bin positions carry appreciable sampling noise.

Brier score depends on their absolute values; a model can therefore rank patients better while placing its probabilities at systematically wrong magnitudes. Fig. 3 shows this be-haviour, all three curves tracking the diagonal loosely with visible excursions. The bin-to-bin volatility partly reflects the small number of records per bin (154 across seven bins), so individual bin positions should not be over-interpreted.

The practical consequence is that model selection depends on intended use: a ranked triage worklist favours XGBoost, whereas a probability compared against an absolute referral threshold is better served by Random Forest's calibrated esti-mates. Reporting accuracy alone would not distinguish these cases.

D. Global Feature Attribution

Fig. 4 reports the five most important features for each model under the mean-absolute-SHAP criterion of (4), and Table III gives the corresponding numerical values. Fig. 5 shows the full per-record attribution distribution for the Ran-dom Forest model, which resolves not only the magnitude of each feature's influence but also its direction.

The three models agree on the two leading predictors—Glucose first and BMI second—a convergence that aligns with clinical understanding of type 2 diabetes pathophysiology. Beyond the second position, agreement deteriorates. Logistic Regression ranks Pregnancies third, ahead of the Diabetes Pedigree Function and Age; both tree ensembles instead rank Age third and place Insulin in their top five, which Logistic Regression does not. Insulin's appearance deserves particular caution, since nearly half of its values are imputed medians (Fig. 1(b)), so an ensemble may be splitting on the artificial spike of identical values rather than on genuine biochemical signal.

The absolute magnitudes in Table III are *not* comparable across models: Logistic Regression and XGBoost attributions are in log-odds units, whereas Random Forest attributions are in probability units, which is why the latter are roughly an order of magnitude smaller. Only the relative ordering within

TABLE II
HELD-OUT TEST PERFORMANCE (n = 154: 100 NON-DIABETIC, 54 DIABETIC)

Model	Accuracy	Non-diabetic (class 0)			Diabetic (class 1)			ROC-AUC	Brier score
		Precision	Recall	F_1	Precision	Recall	F_1		
Logistic Regression	0.7078	0.75	0.82	0.78	0.60	0.50	0.55	0.8130	0.1754
Random Forest	0.7468	0.85	0.74	0.79	0.61	**0.76**	**0.68**	0.8239	**0.1646**
XGBoost	**0.7597**	0.83	0.79	**0.81**	**0.64**	0.70	0.67	**0.8374**	0.1656

TABLE III
GLOBAL MEAN ABSOLUTE SHAP VALUES (TOP FIVE PER MODEL)

Rank	Feature	Mean $\|\phi_j\|$
Logistic Regression		
1	Glucose	1.0155
2	BMI	0.5664
3	Pregnancies	0.3413
4	Diabetes Pedigree Func.	0.1740
5	Age	0.1263
Random Forest		
1	Glucose	0.1363
2	BMI	0.0839
3	Age	0.0591
4	Insulin	0.0408
5	Diabetes Pedigree Func.	0.0318
XGBoost		
1	Glucose	1.3060
2	BMI	0.8037
3	Age	0.5486
4	Diabetes Pedigree Func.	0.4773
5	Insulin	0.3079

each column is meaningful—a point that bears directly on the FCI, as discussed in Section V.
Fig. 5 adds directional information that magnitude rankings omit: high Glucose values push predictions toward the dia-betic class while low values push them away, and the same monotone pattern holds for BMI and Age. This monotonicity was not imposed as a constraint and provides a form of face validity that an importance ranking alone cannot supply.

E. Feature Consistency and Composite Ranking

Table IV reports, for every feature, the mean cross-model importance μ_j, the coefficient of variation CV_j, the Feature Consistency Index of (7), and the composite score S_j of (9). Fig. 6 visualises the ranking.
Glucose leads on every criterion: it has by far the largest mean importance (μ = 0.819), the highest consistency (FCI = 0.622), and consequently the highest composite score (S = 0.849). BMI is a clear second (S = 0.580) with essentially equal consistency (FCI = 0.618). The gap between these two and the remainder is substantial: the third-ranked feature scores 0.354, less than half of Glucose's score.
The middle of the ranking is closely contested, and its ordering is driven by consistency rather than importance. Age, the Diabetes Pedigree Function, and Pregnancies occupy positions three through five with composite scores of 0.354, 0.348, and 0.345—a spread below 0.01—although their mean importances differ more (μ = 0.245, 0.228, 0.198), because Pregnancies compensates for lower importance with markedly higher consistency (FCI = 0.602 against Age's 0.530). This

TABLE IV
FEATURE CONSISTENCY INDEX AND COMPOSITE RANKING

Rank	Feature	μ_j	CV_j	FCI_j	S_j
1	Glucose	0.8193	0.6069	0.6223	**0.8489**
2	BMI	0.4847	0.6179	0.6181	0.5804
3	Age	0.2447	0.8855	0.5304	0.3539
4	Diabetes Pedigree Func.	0.2277	0.8160	0.5507	0.3484
5	Pregnancies	0.1981	0.6623	0.6016	0.3452
6	Insulin	0.1318	0.9453	0.5141	0.2573
7	Blood Pressure	0.0962	1.0934	0.4777	0.2144
8	Skin Thickness	0.0670	1.0137	0.4966	0.1986

is the FCI functioning as designed: a moderately important feature all models agree upon is elevated over a slightly more important one they dispute. Given the narrow margins, the three are best regarded as a tied band. At the bottom, Blood Pressure and Skin Thickness are both weakly attributed and the least consistent, with the two highest coefficients of variation (1.093 and 1.014, both exceeding unity). They are therefore poor candidates for a reduced feature set on both grounds simultaneously. Insulin occupies an intermediate position ($S = 0.257$) that should be read in light of its 48.7% imputation rate.

F. Patient-Level Explanation

Cohort-level importance describes average behaviour and does not necessarily describe any particular patient. To ex-amine this, the framework was applied to a synthetic but clinically plausible record: a 60-year-old woman with six prior pregnancies, plasma glucose 110 mg/dL, diastolic blood pressure 85 mm Hg, skin-fold thickness 40 mm, serum insulin 200 μU/mL, BMI 38 kg/m^2, and a diabetes pedigree function of 1.2. This patient's glucose level is near the cohort median, but her BMI, age, and pedigree function are all substantially elevated. All three models classify her as at elevated risk, with predicted probabilities of 0.5639 (Logistic Regression), 0.6576 (Random Forest), and 0.7635 (XGBoost). Table V reports the per-model local attributions and the resulting patient-specific consistency ranking, and Fig. 7 decomposes the XGBoost prediction. The ranking for this patient differs materially from the cohort ranking of Table IV. The Diabetes Pedigree Function, only fourth at the cohort level, is the dominant local factor ($S = 0.820$), followed by Glucose (0.516) and BMI (0.513). This inversion is exactly what should be expected and is the reason local explanation is necessary: because her glucose is unremarkable, it cannot be what distinguishes her from the average patient, whereas her strongly elevated pedigree function and BMI can. The waterfall plot in Fig. 7 makes

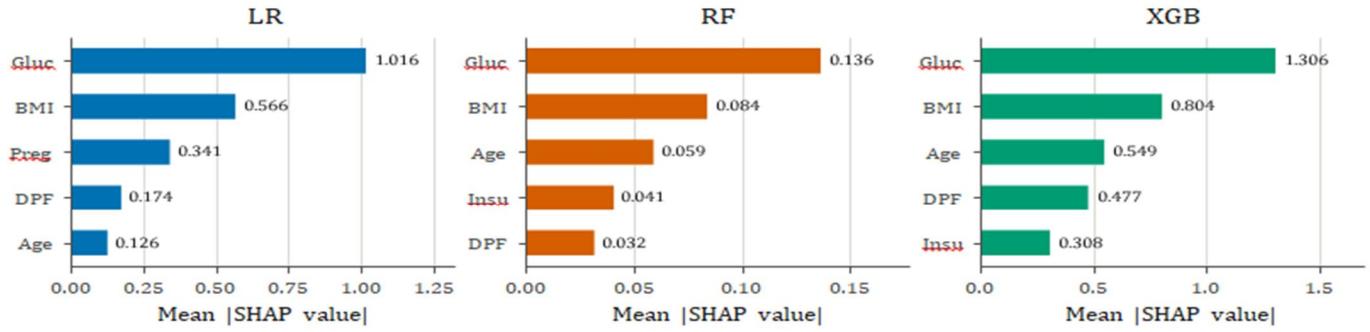


Fig. 4. Top five features by mean absolute SHAP value for each model. All three agree on Glucose and BMI in the first two positions; rankings diverge thereafter. Magnitudes are *not* comparable across panels—Random Forest attributions are in probability units, the other two in log-odds units.

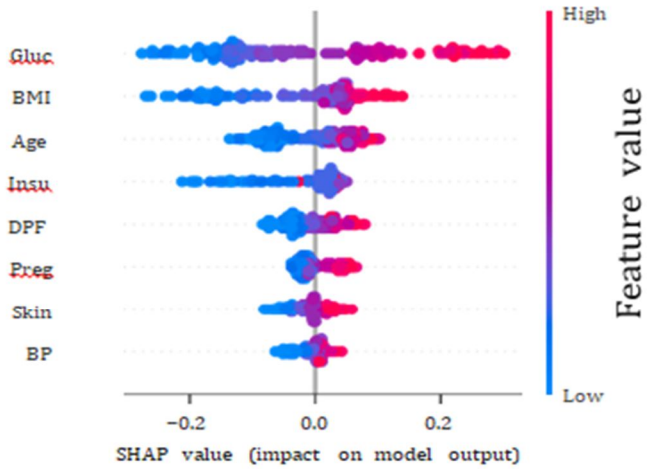


Fig. 5. Per-record SHAP attributions for the Random Forest model on the test set. Each point is one patient; horizontal position gives the signed contribution and colour encodes the feature value. High Glucose, BMI, and Age values consistently increase predicted risk.

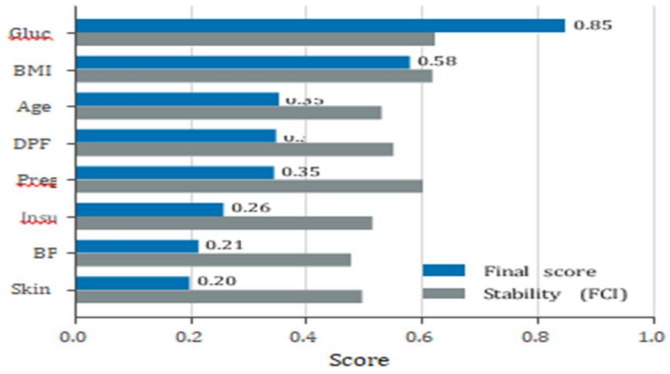


Fig. 6. Composite score S_j and the underlying consistency component FCI_j for all eight predictors. Glucose and BMI dominate on both criteria; Blood Pressure and Skin Thickness are simultaneously the least important and the least consistently attributed.

the mechanism explicit—the pedigree function contributes the largest positive push toward the diabetic class, while her near-median glucose value actually contributes *negatively*, pulling the prediction down relative to the cohort baseline.

TABLE V

PATIENT-LEVEL SHAP ATTRIBUTIONS AND CONSISTENCY RANKING

Rank	Feature	LR	RF	XGB	FCI_j	S_j
1	Diabetes Pedigree Func.	0.5177	0.0754	1.3571	0.5502	**0.8201**
2	Glucose	0.3795	0.0858	0.4201	0.6646	0.5162
3	BMI	0.5521	0.0544	0.3399	0.6074	0.5134
4	Age	0.3100	0.0518	0.2427	0.6482	0.4175
5	Pregnancies	0.2018	0.0122	0.2936	0.5909	0.3628
6	Insulin	0.0516	0.0380	0.0330	0.8388	0.3355
7	Skin Thickness	0.0359	0.0234	0.1154	0.5883	0.2524
8	Blood Pressure	0.0451	0.0151	0.1540	0.5447	0.2479

$f(x) = 1.621$

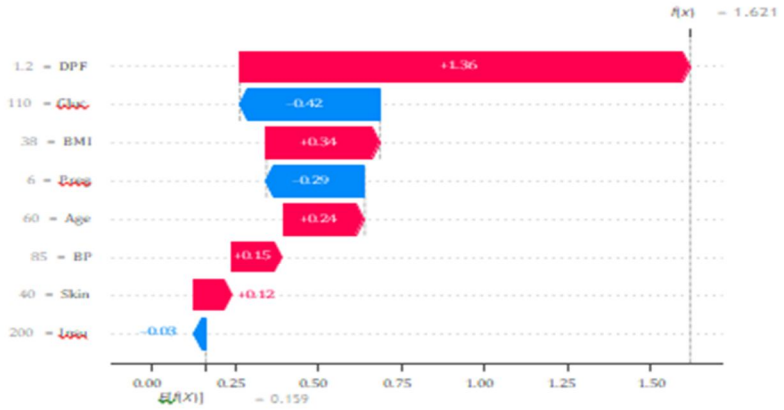


Fig. 7. SHAP waterfall decomposition of the XGBoost prediction for the patient of Section IV-F. The Diabetes Pedigree Function supplies the largest positive contribution, while the patient's near-median Glucose value contributes negatively—inverting the cohort-level ordering of Table IV.

For a clinician, this is the actionable output: this patient's risk is driven predominantly by hereditary predisposition and adiposity rather than by current glycaemic status, which points toward weight management and structured surveillance rather than toward immediate glycaemic intervention.

G. Robustness to Input Perturbation

Fig. 8 reports the effect of the 5% Gaussian perturbation of (10) on the same patient record. Logistic Regression shifts from 0.5639 to 0.5737 ($\Delta = 0.0097$) and Random Forest from 0.6576 to 0.6498 ($\Delta = 0.0078$), whereas XGBoost shifts from 0.7635 to 0.8022 ($\Delta = 0.0386$)—approximately four to five times the sensitivity of the other two models.

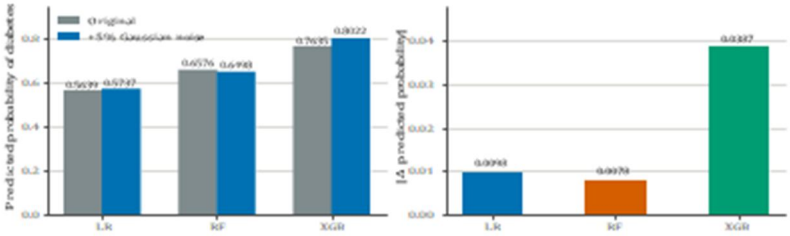


Fig. 8. Robustness to a 5% Gaussian perturbation of the patient record. Left: predicted probability before and after perturbation. Right: absolute change. XGBoost, the most accurate model, is roughly four to five times more sensitive than the other two.

The ordering follows from the models' inductive biases. Logistic Regression is linear in the inputs, so its response to a small perturbation is bounded by the corresponding coefficient and necessarily smooth. Random Forest averages 200 independently grown trees, which suppresses the individ-ual threshold crossings a perturbation induces in any single tree. Boosting provides neither protection: successive trees fit residuals and therefore construct finer partitions precisely where the ensemble is currently mistaken, producing a sharper, more locally responsive decision surface. Critically, this ordering is the *inverse* of the accuracy order-ing in Table II. The most accurate model is the least stable, and the least accurate is among the most stable. A purely accuracy-driven model selection would choose XGBoost without reg-istering that its outputs move most under measurement-scale variation; a reliability-aware evaluation surfaces this trade-off explicitly and allows it to be weighed against the accuracy gain. It should be stressed that this analysis rests on a single record and a single noise realisation, and its limitations are addressed in Section VI.

V. DISCUSSION

A. Accuracy, Reliability, and Stability Are Distinct Axes

The single most consequential result of this study is that the three models rank differently depending on the axis of evaluation. XGBoost is best on discrimination (accuracy 0.7597, AUC 0.8374); Random Forest is best on probabilistic reliability (Brier 0.1646) and on positive-class recall (0.76); and Random Forest and Logistic Regression are jointly best on robustness, being four to five times less sensitive to input perturbation than XGBoost. No single model dominates.

Had this study reported accuracy alone, as much of the applied literature does, it would simply have recommended XGBoost for deployment—a conclusion correct on its own terms and materially incomplete. The appropriate choice de-pends on how the output is consumed: a ranked worklist favours the best discrimination, an absolute referral threshold favours the best-calibrated probabilities, and a setting with noisy measurements favours the most stable model. Making these axes explicit converts an invisible modelling decision into a reviewable one.

B. What the Feature Consistency Index Adds

The FCI addresses a question that single-model importance analysis cannot pose: how much of the reported explanation is a property of the data, and how much is a property of the model that happened to be selected? The results give a graded answer. Glucose and BMI are stable across all three model families and can be reported with confidence as genuine drivers. Age, the Diabetes Pedigree Function, and Pregnancies form a contested middle band whose internal ordering is not robust—any of the three could be promoted by a different model choice, and the sub-0.01 spread in their composite scores should be read as a tie rather than a ranking. Blood Pressure and Skin Thickness are consistently unimportant, which is itself a stable and useful finding.

Reporting this graded structure is more honest than report-ing a single model's ranked list, which conveys a precision the evidence does not support.

Two properties of the current formulation warrant scrutiny. First, the coefficient of variation in (7) is computed over raw SHAP magnitudes that are not on a common scale across models (Table III). A substantial part of the measured dispersion is therefore attributable to this unit mismatch rather than to genuine disagreement, which depresses all FCI values toward the middle of their range—no feature in Table IV exceeds 0.63, not even Glucose, on which all three models plainly agree. Normalising each model's importance vector to unit sum before computing CV_j would isolate relative disagreement from scale, and we regard this as the most important refinement of the measure. The current formulation remains informative for *relative* comparison between features, since the mismatch is common to all of them, but its absolute values should not be read as agreement percentages.

Second, the weighting $\alpha = 0.6$ in (9) is a design choice rather than an estimated quantity. The top two positions are insensitive to it—Glucose and BMI lead on both components separately, hence for any $\alpha \in [0, 1]$—but the contested middle band is not. Reporting the ranking under several values of α, or the two components separately as in Fig. 6, is preferable to presenting one composite ordering as definitive.

C. Global Versus Local Explanation

The patient-level analysis of Section IV-F demonstrates con-cretely that the cohort ranking does not transfer to individuals. For the examined patient, the Diabetes Pedigree Function—fourth at cohort level—became the dominant factor, while Glucose contributed negatively because her value was near the cohort median. Presenting this patient with the cohort-level explanation "your risk is driven mainly by glucose and BMI" would have been actively misleading with respect to the first of those factors.

This has a direct design implication for any clinical decision-support interface: the explanation shown to the clin-ician must be computed for the patient in front of them, not retrieved from a global importance table. The cost of TreeSHAP for a single record is negligible, so there is no practical reason to do otherwise.

D. Clinical Plausibility

The framework's outputs are consistent with established type 2 diabetes pathophysiology. Glucose dominating the attribution is expected, though partly definitional given that the outcome label derives from glucose-based diagnostic cri-teria; BMI ranking second reflects the well-established role of adiposity in insulin resistance; and the monotone directional relationships visible in Fig. 5—higher glucose, BMI, and age increasing predicted risk—match clinical expectation without having been imposed as constraints.

Blood Pressure ranking last is superficially surprising given the known association between hypertension and diabetes, but it is coherent here: that association is largely mediated by shared risk factors such as adiposity and age, both already rep-resented among the predictors, so blood pressure contributes little additional independent information. Insulin, conversely, warrants more caution than its middle ranking suggests: with 48.7% of its values imputed to a single median, the feature is close to a constant with a minority of informative values, and its intermediate importance should not be read as evidence of genuine biochemical signal.

VI. LIMITATIONS

Several limitations bound the conclusions of this study and should be addressed before any of its findings are treated as clinically actionable.

Dataset scope. The dataset comprises 768 records from a single, demographically narrow population—female patients of Pima Indian heritage aged 21 and above—with an excep-tionally high recorded prevalence of type 2 diabetes. Both the base rate and the learned feature–outcome relationships may therefore not transfer to other populations, to men, or to younger patients. No external validation cohort was available, and external validation is the single most important prerequisite for clinical use.

Imputation. Median imputation was applied to five pre-dictors, two of which—Insulin (48.7%) and Skin Thickness (29.6%)—are affected in a very large share of records. Median imputation assumes data are missing at random, an assumption unlikely to hold here, since a missing insulin measurement plausibly correlates with the clinical circumstances of the visit. It also artificially reduces variance and creates a spike of identical values that tree-based models can split on spuri-ously. Multiple imputation, or explicit missingness indicators allowing the models to use the fact of absence as infor-mation, would be methodologically preferable. Additionally, imputation medians were computed over the full cohort before partitioning, introducing a mild optimistic bias; a fully nested implementation would fit the imputer within the training fold only.

Robustness analysis. The perturbation study rests on a single synthetic record and a single noise realisation at one noise level, and is therefore an illustration of a sensitivity difference rather than an estimate of one. The observed ordering is consistent with the models' inductive biases and with the wider literature on the fragility of boosted models, but establishing it properly requires a Monte Carlo study over the full test set with many draws at several values of λ, reporting the distribution of $\Delta^{(m)}$ rather than a point value. This is the most immediate extension of the present work.

Statistical uncertainty. All results derive from a single strat-ified split with 154 test records, without confidence intervals or significance tests. Differences of the magnitude separating Random Forest and XGBoost on the Brier score (0.1646 versus 0.1656) lie well within the sampling variability of a test set this small; repeated stratified cross-validation would be required to establish which observed differences are real.

Scale dependence of the FCI. As discussed in Section V, the current FCI formulation mixes genuine cross-model disagree-ment with differences in attribution units. Its relative ordering of features is informative; its absolute values are not directly interpretable as agreement scores.

Post-hoc explanation caveats. SHAP's standard estimators assume feature independence, violated by the correlated pre-dictors in Fig. 2, so attribution may be distributed arbitrarily within correlated groups such as BMI and Skin Thickness [20]. More fundamentally, SHAP explains the model, not the dis-ease: high attribution identifies a feature the model relies upon, which is not the same as a causal risk factor, and no conclusion about intervention should be drawn from attribution magnitude alone [19].

VII. CONCLUSION AND FUTURE WORK

This paper presented a reliability-aware and interpretable machine learning framework for diabetes prediction from structured clinical data, evaluating Logistic Regression, Ran-dom Forest, and XGBoost along four coordinate axes: dis-crimination, probabilistic reliability, interpretability, and ro-bustness.

The results show that these axes are genuinely distinct. XGBoost achieved the best discrimination (accuracy 0.7597, ROC-AUC 0.8374), Random Forest the best-calibrated proba-bilities (Brier score 0.1646) together with the highest recall on the diabetic class (0.76), and Logistic Regression and Random Forest were markedly more robust to input perturbation than XGBoost, whose predicted probability moved four to five times further under a 5% measurement-scale perturbation. Model selection therefore cannot be settled by accuracy alone, and the appropriate choice depends on how the model's output will be consumed clinically.

The proposed Feature Consistency Index quantified the degree to which SHAP attributions agree across model fam-ilies, converting a normally invisible source of explanatory uncertainty into a reported quantity. It identified Glucose and BMI as both the most influential and the most consistently attributed predictors; Age, the Diabetes Pedigree Function, and Pregnancies as a contested middle band whose internal ordering is not robust to model choice; and Blood Pressure and Skin Thickness as consistently uninformative. Patient-level analysis further showed that the cohort ranking does not transfer to individuals, supporting per-patient explanation in any deployed interface.

Future work follows directly from the limitations of Sec-tion VI. The priorities are: external validation on an indepen-dent and demographically broader cohort; replacement of the single split with repeated stratified cross-validation reporting confidence intervals; a full Monte Carlo robustness study across the test set at multiple noise levels; scale-normalisation of the FCI so that its absolute values become interpretable, together with a comparison against rank-based agreement statistics such as Kendall's W; multiple imputation or explicit missingness indicators in place of median imputation; and the addition of post-hoc recalibration by Platt scaling or isotonic regression, which would allow discrimination and calibration to be optimised separately rather than traded off.

VIII. REPRODUCIBILITY

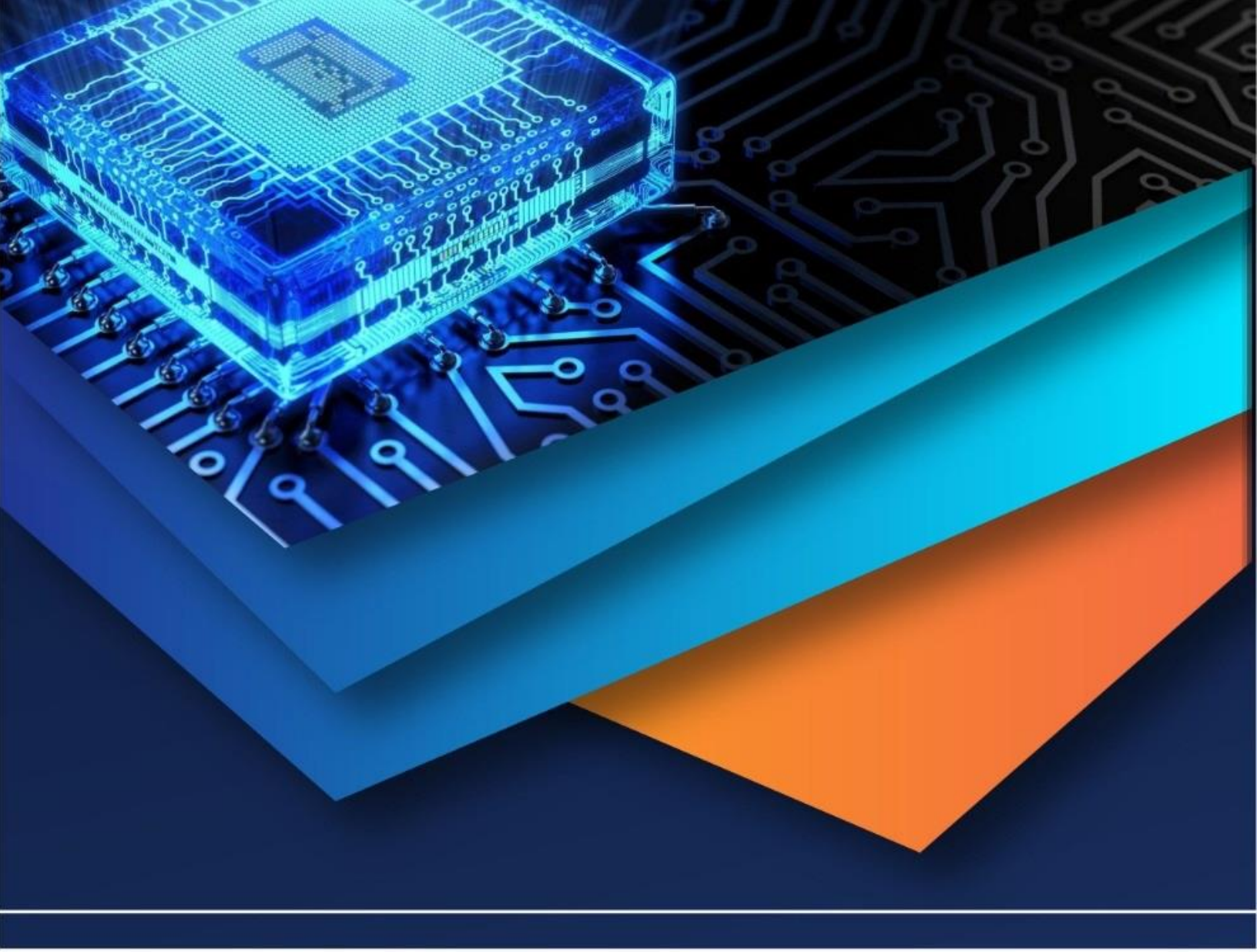
All experiments use a fixed random seed (42) for parti-tioning and model fitting. Models were implemented with scikit-learn [30] and XGBoost [29], with attributions computed using the shap library [15]. Readers reproducing these results should note that the exact test-set metrics of gradient-boosted models are sensitive to the XGBoost build and platform, because row and column subsampling draw from a pseudo-random stream whose implementation has changed across releases; the Logistic Regression and Random Forest results reproduce exactly across platforms, whereas XGBoost metrics may vary in the second decimal place. Figures were regen-erated from the same pipeline and, for XGBoost, reflect a re-execution of the reported configuration.

IX. ACKNOWLEDGMENT

The authors thank [*acknowledgements to be completed*].

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.
[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.
[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Frontiers in Genetics, vol. 9, art. 515, 2018.
[4] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," Health Information Science and Systems, vol. 8, art. 7, 2020.
[5] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proc. Annual Symposium on Computer Application in Medical Care, 1988, pp. 261–265.
[6] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," BMC Medicine, vol. 17, art. 230, 2019.
[7] E. W. Steyerberg et al., "Assessing the performance of prediction models: a framework for traditional and novel measures," Epidemiology, vol. 21, no. 1, pp. 128–138, 2010.
[8] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individ-ual prognosis or diagnosis (TRIPOD): the TRIPOD statement," BMJ, vol. 350, art. g7594, 2015.
[9] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.
[10] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd International Conference on Machine Learning (ICML), 2005, pp. 625–632.
[11] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, A. J. Smola et al., Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
[12] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. 8th ACM SIGKDD Interna-tional Conference on Knowledge Discovery and Data Mining, 2002, pp. 694–699
[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.
[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.
[15] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
[16] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton University Press, 1953, pp. 307–317.
[17] E. Šˇtrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," Knowledge and In-formation Systems, vol. 41, no. 3, pp. 647–665, 2014.
[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
[20] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: more accurate approximations to Shapley values," Artificial Intelligence, vol. 298, art. 103502, 2021.
[21] M. Sundararajan and A. Najmi, "The many Shapley values for model ex-planation," in Proc. 37th International Conference on Machine Learning (ICML), 2020, pp. 9269–9278.
[22] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2020, pp. 180–186.
[23] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in Proc. AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3681–3688.
[24] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretabil-ity methods," in Proc. ICML Workshop on Human Interpretability in Machine Learning (WHI), 2018.
[25] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," Knowledge and Infor-mation Systems, vol. 12, no. 1, pp. 95–116, 2007.
[26] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," Journal of Machine Learning Research, vol. 18, no. 174, pp. 1–54, 2018.
[27] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.
[28] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5– 32, 2001.
[29] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[30] F. Pedregosa et al., "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY