IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5522∼5533
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5522-5533

Explainable artificial intelligence for enhanced analysis of misclassified diabetes predictions

Deepthi Angara, Jhansi Yellapu
Department of Computer Science and Engineering, GITAM School of Technology (Deemed to be University), Visakhapatnam, India

Article Info

Article history:

Received Nov 17, 2025
Revised Aug 17, 2026
Accepted Sep 20, 2026

Keywords:

BRFSS2015 dataset
Diabetes prediction
Diverse counterfactual explanations
Explainable artificial intelligence
Interpretability
Local interpretable model-agnostic explanations
Post-hoc explainability

ABSTRACT

Misclassified cases in diabetes prediction, i.e., false positives (FP) and false negatives (FN), carry clinical risks. Using the behavioral risk factor surveillance system (BRFSS)-2015 dataset and a tuned gradient-boosting benchmark, this study identified FP/FN cohorts. This study introduces local interpretability with counterfactual explanations (LICE), an enhanced hybrid explainable artificial intelligence (XAI) framework centered on these errors. This study validated reliability with local interpretable model-agnostic explanations (LIME) fidelity and diverse counterfactual explanations (DiCE) cosine similarity, Spearman correlation, and Jaccard overlap. This study then applied LICE-guided updates, including feature-influence modification and out-of-fold threshold calibration, and evaluated the resulting model on the held-out test set. LIME resulted in a fidelity of 0.9187 across FP/FN and highlighted influential features as over-resilient and underutilized. DiCE resulted in cosine and Spearman $>$ 0.96; Jaccard of 0.38 (FP) and 0.28 (FN), indicating diverse counterfactuals with preserved proximity. Compared with the tuned benchmark (area under the curve (AUC) 0.8310), the LICE-enhanced gradient boosting decision tree (EnhancedGBDT) maintained AUC 0.8309 (Δ -0.0001) and shifted towards sensitivity: recall = 0.8203 (+0.02), precision = 0.7370, and F1 = 0.7764. Confusion matrix changes confirm FN -143 and FP +119. Relative to the untuned baseline (AUC 0.8304), AUC improved, and recall gains held, indicating that error-focused explanations enhance reliability without sacrificing overall discrimination.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Deepthi Angara
Department of Computer Science and Engineering, GITAM School of Technology (Deemed to be University)
Visakhapatnam, Andhra Pradesh, India
Email: dangara@gitam.in

1. INTRODUCTION

Machine learning (ML) models have strong decision-making but lack transparency. Bias, data shifts, and generative hallucinations can distort model outcomes and lack transparency [1]. Opacity-aware management, combined with explanation-driven actions, boosts the reliability and trustworthiness of the ML model [2], [3]. Explainable artificial intelligence (XAI) bridges the gap between the decision-making of ML models and the transparency of their decisions.

Longo *et al.* [4] have deployed XAI in research, development, and deployment for the past decade. According to the XAI 2.0 manifesto, interdisciplinary collaboration should advance transparency on the practical limitations of ML models. XAI is essential in prioritizing trust, ethics, legal demands, user viewpoints,

and application orientation [5]. Black-box models are highly accurate but also challenging to comprehend and trust, especially in sensitive domains like healthcare [6], [7]. Mersha *et al*. [8] identify healthcare as a domain where opacity is especially problematic. According to the World Health Organization (WHO), the number of diabetes cases has approximately doubled from 7% to 14% worldwide between 1990 and 2022 [9]. Majcherek *et al*. [10] addressed chronic hyperglycemia resulting from impaired insulin function, the occurrence of comorbidities such as heart disease in individuals with diabetes, as well as diet-related and lifestyle-related type 2 diabetes (T2D). Similarly, Galicia-Garcia *et al.* [11] in their paper addressed insulin resistance as the root cause of cardiovascular disease. Fregoso-Aparicio *et al.* [12] stated in their paper that hundreds of millions are affected globally, with rapid growth expected, and a significant fraction of cases remain undiagnosed. They also noted that the diversity of algorithms makes it hard to select the best models; at the same time, the lack of feature transparency reduces interpretability.

Based on the attention required for undiagnosed diabetes, which may lead to other comorbidities, and considering the ML models' strong decision-making capabilities that involve integrating transparency, this study proposes integrating an ML model with a hybrid XAI framework called local interpretability with counterfactual explanations (LICE). LICE is a hybrid framework that integrates local interpretable model-agnostic explanations (LIME), which works on instance-level data points, while diverse counterfactual explanations (DiCE) generates feasible counterfactuals to investigate the decision boundary [13], [14].

This study separated false case (FC) cohorts, i.e., false positives (FP) and false negatives (FN), which carry different clinical risks post-modeling. FPs may lead to unnecessary workups, while FNs can delay care. This study measured fidelity [15] and stability [16] to assess the LICE framework on FC cohorts. Niño-Adan *et al.* [17] in their paper addressed the fact that the weights of the features vary by instance, region, or class. Considering the LICE-informed feature weights, this study decided to reweight a set of features suggested by LICE to improve undiagnosed FN that delays the necessary care. This way, this study focused on the interpretability that impacts safety, and complements human-centered, responsible artificial intelligence (AI) [18], [19]. Unlike traditional XAI approaches, which primarily analyze correctly classified instances, the novelty of LICE lies in its focus on misclassifications to uncover deeper model behavior. LIME is used for identifying FP features that overly contribute to positive predictions and FN features that are insufficiently contributing to risk factorization. Counterfactual explanations provide the minimum amount of change that will either correct a prediction or find a missed case; they highlight how sensitive a model is to both underweighting and overweighting different factors. Through the systematic analysis of FC cohorts and by using the explanation-based insights to improve the model, LICE closes the gap between model improvement and model interpretability, allowing for explanation-guided interventions.

Key contributions of our LICE framework are as follows:

- Novel hybrid XAI framework–LICE is introduced, which integrates LIME (local attribution) and DiCE (counterfactuals) around a trained ML model to provide transparency.
- LICE-guided error analysis of FP and FN cases–misclassifications in the ML model that identify over-resilient and underutilized features are diagnosed.
- LICE-guided model enhancement–insights are translated into an interpretable adjustment: mild feature reweighting, threshold calibration, and simple interaction terms, and the precision-recall trade-off in clinical relevance is reported.
- Quantitative evaluation of explanations and prediction–fidelity and stability of explanations are addressed, and Cohen's kappa and Matthews correlation coefficient (MCC) are used as robust model metrics.

2. RESEARCH GAP AND PROPOSED METHOD

In diabetes prediction, work largely optimizes aggregate metrics (accuracy/area under the curve (AUC)) and, when XAI is applied, utilizes post-hoc feature importance plots (e.g., Shapley additive explanations (SHAP)) rather than cohort-level analysis of misclassifications (FP/FN) or explainability-guided model enhancements. In explainability, LIME provides instance-level local explanations for black-box models [13]. DiCE generates diverse counterfactuals that probe local decision boundaries [14]. Evaluating explanation quality remains active, including fidelity and stability considerations relevant to clinical reliability [15], [16]. Across these threads, the diabetes prediction still lacks in: i) FP/FN-focused explainability, ii) diagnostic use of counterfactuals to collect facts, and iii) linking these diagnostics to principled influence modulation and threshold calibration. Table 1 shows the comparative analysis of existing work.

This study addresses this gap by focusing on FP/FN cohorts in the behavioral risk factor surveillance system (BRFSS)-2015. LIME was used to characterize the local reasons for each error and DiCE to identify the decision boundary with diversity, feasibility-constrained counterfactuals as diagnostic instruments. These reasons were then summarized to inform influential decisions, decrease the weight of over-resilience, and increase the weight of underutilized features. This was followed by Youden's J-based threshold calibration and finally evaluating reliability via fidelity/stability metrics on the error subsets [13]–[16].

Table 1. Comparison of existing work

Paper	Dataset(s)	Model(s)	Metrics	Key results	XAI used?	Summary
Zhang *et al.* [20]	PIMA, BRFSS-2015 and BIT Mesra	Custom BPNN (+ balancing)	Accuracy, sensitivity, and specificity	Accuracy: PIMA 0.89 and BRFSS 0.75	None	Cross-dataset portability and undersampling influences comparability
Mansouri *et al.* [21]	PIMA	K-nearest neighbor (KNN)	Accuracy, precision, recall, and F1	Accuracy 0.76, precision 0.80, recall 0.85, and F1 0.83	None	Reproducible pipeline and single-model demo
Majcherek *et al.* [10]	BRFSS-2015 (∼441K) + others	18 models (random forest (RF), extreme gradient boosting (XGB), and light gradient-boosting machine (LGBM))	Accuracy and AUC	Accuracy $>$ 0.90	SHAP (qualitative)	Highlights body mass index (BMI)/age and qualitative interpretation only
Halias *et al.* [22]	BRFSS-2015 (∼70K)	Nine prediction models and logistic regression (LR) with optimal selection	Accuracy	Accuracy 0.7561	None	Data mining techniques for T2D prediction
Chang *et al.* [23]	BRFSS-2015 (∼70K)	LR, support vector machine (SVM), KNN, naive Bayes (NB), decision trees (DT), and RF	Accuracy, precision, recall, and F1	RF Acc 0.82	None	Practical replication guide
Jose *et al.* [24]	BRFSS-2015	LightGBM, gradient boosting decision tree (GBDT), AdaBoost, LR, RF, and SVM	Accuracy, precision, recall, F1, AUC, Kappa, and MCC	LightGBM: accuracy 0.7536, AUC 0.8302, and GBDT: AUC 0.8309	SHAP and LIME	Strong discrimination and limited XAI depth

3. METHODOLOGY

The Centers for Disease Control and Prevention (CDC) conducts the annual BRFSS [25], [26]. The proposed method used the CDC dataset `diabetes_binary_5050split_health_indicators_BRFSS2015.csv` [27], which is balanced for the diabetes outcome and contains 70,692 responses. BRFSS-2015 is a large, population-level survey that contains diverse demographic, behavioral, and health-indicator variables relevant to chronic disease risk, including diabetes, cardiovascular conditions, and lifestyle factors. Moreover, the dataset is self-reported, which introduces realistic noise and imbalance, making it a suitable testbed for studying explainability models. Compared to more controlled or smaller datasets, this dataset supports robust local explanations for different cohorts and explanation-guided model interventions, aligning closely with the goals of interpretable and actionable AI in public health analytics.

Before beginning data processing and preparation, 2% duplicate records were removed as part of the preprocessing steps to create an almost perfectly balanced dataset comprising 69,057 instances. Outliers in continuous variables, including BMI, mental health, and physical health, were then verified and imputed to mitigate distortion [28]. Finally, the dataset was considered for predictive modeling and interpretability.

Jose *et al*. [24] was adopted as the base paper, as they trained multiple classifiers on the BRFSS-2015 dataset. To ensure compatibility, the evaluation metrics were aligned with their reporting of the GBDT classifier on diabetes prediction using the BRFSS-2015 data. To the best of our knowledge, this is the first study to align a reproduction with de-duplication. LICE is proposed as a hybrid XAI framework that combines LIME [13] for

instance-level attribution with DiCE [14] for actionable recourse, going beyond accuracy. LICE leverages XAI to diagnose misclassified cases and inform targeted refinements for safe and clinically interpretable results.

Figure 1 illustrates the proposed architecture of LICE. Algorithm 1 outlines the LICE insight generation process, which begins with a trained model and FP/FN cohorts, followed by computing LIME attributions, generating DiCE counterfactuals for misclassified instances, and assessing fidelity and stability. Finally, cohort-level guidance is aggregated into safe, interpretable use.

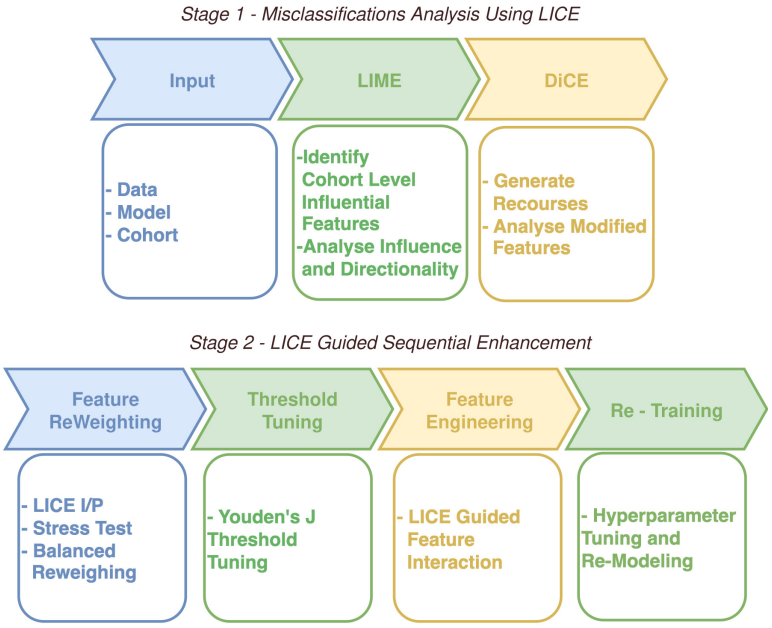


Figure 1. LICE misclassifications analysis and retraining architecture

Algorithm 1: LICE framework insights generation for misclassifications

Data: Trained benchmark model M_0; dataset $D = (\mathbf{X}, \mathbf{y})$; cohort partition $C \in \{\text{FP}, \text{FN}\}$.
Result: Cohort-level insights $\mathcal{I}$.

Step 1: Local explanation (LIME)
foreach $x \in C$ **do**
  Fit a locally faithful surrogate g_x on perturbed samples around x and estimate local feature weights $\mathbf{w}(x)$
  Extract top-k influential features $\mathcal{F}_1(x) = \{f_1, \ldots, f_k\}$ (with sign $\pm$)
Step 2: Actionable recourse (DiCE)
foreach $x \in C$ **do**
  Pass $\mathcal{F}_1(x)$ to DiCE for targeted counterfactual generation
  Generate a diverse set $\{x'_1, \ldots, x'_m\}$ such that $M(x'_j)$ flips the decision with minimal change, subject to feasibility and immutability constraints
  Identify DiCE-influential features $\mathcal{F}_2(x) = \{f_1, \ldots, f_\ell\}$
Step 3: Stability audit
Compute stability across the generated $\{x'_j\}$ using: cosine similarity, Spearman rank correlation, and Jaccard overlap
Step 4: Aggregate cohort-level guide to model retraining
Calibrated feature reweighting/engineering;
threshold tuning using out-of-fold training predictions and Youden's J;
hyperparameter search;
re-evaluate and iterate.

Algorithm 2 presents the LICE-guided model enhancement procedure, which implements the explanation-driven interventions summarized in Table 2. Feature influence was adjusted through the reduction and increase of selected predictors, guided by LIME feature importance and supported by patterns observed in DiCE-generated counterfactual recourse. Predictors that showed excessive influence on model decisions were reduced, whereas underutilized predictors were strengthened to encourage a more balanced contribution across features. This controlled modification of feature influence served as a practical proxy for reweighting feature importance without altering the underlying categorical semantics or data structure.

Algorithm 2: LICE-guided model enhancement

Data: Tuned GBDT M_0; insights $\mathcal{I}$ from Algorithm 1; scaling presets $S_{\text{mild}}, S_{\text{aggr}}$; interaction set $\{HighBP \times HighChol\}$; risk-amplifier $HeartDiseaseorAttack_amp = 1.5$; threshold T.
Result: Updated model $M^\star$, operating threshold $\tau^\star$, and accepted interventions.

Acceptance gates (apply after each candidate):
  $\text{AUC} \geq \text{AUC}(M_0) - \varepsilon$ (AUC floor) and $\%\Delta\text{FP} \leq \Delta_{\max}$ (e.g., 6%);
  Stability maintained: cosine ≥ 0.96, Spearman ≥ 0.86, median Jaccard ≥ 0.25.

Step 1: Feature reweighting
  (a) Reactive reweighting — stress test
    Apply a strong reduction to over-resilient predictors and strengthen underutilized predictors.
    Retrain GBDT without hyperparameter tuning.
    Evaluate at τ_0; retain the candidate if gates are met, else reject and proceed.
  (b) Balanced reweighting
    Apply controlled reduction/increase to the same features with small perturbations.
    Retrain without hyperparameter tuning; retain if stability & performance criteria are met.

Step 2: Threshold tuning (Youden's J)
Using cross-validated out-of-fold predictions from the training split, choose $\tau = \arg\max_\tau J(\tau)$ under the FP bound (no retraining).
Retain τ if acceptance gates are satisfied.

Step 3: Feature engineering
Add interaction and apply risk amplifier.
Retrain without hyperparameter tuning.
Re-optimize τ using out-of-fold training predictions and Youden's J; fix the selected threshold before test-set evaluation and apply the acceptance gates.

Step 4: FN-prioritized weighting with hyperparameter tuning
Run GBDT tuning with AUC-floor constraint over
  {`n_estimators, max_depth, learning_rate, subsample, min_samples_leaf`}
Increase FN weights, re-optimize τ by maximizing Youden's J on the OOF predictions; apply gates.
return $M^\star$, $\tau^\star$

Classifier configurations were evaluated on the data using models that included GBDT, SVM, LR, and a stacked ensemble combining LR, SVM, and GBDT. The data was partitioned using a stratified 80/20 train-test split with a seed of 42. Given the need for post hoc local explanations (LIME) and actionable counterfactuals (DiCE), GBDT was selected as the reference classifier for LICE analysis.

GBDT has feature-level reasoning, calibrated thresholds, and provides a means to generate stable probabilities, which are necessary for optimizing Youden's J statistic [29], [30]. GBDT was further tuned with RandomizedSearchCV using 3-fold cross-validation, focused on generalization and stability, and designated as the TunedGBDT benchmark model for LICE and further analyses and interventions. Misclassifications were systematically diagnosed. FP and FN cohorts were defined on the test split at the baseline threshold.

For each FP and FN instance, LIME fits a locally weighted linear surrogate and retains the top-k features, with a default k = 5. DiCE generates diverse counterfactuals per instance, with a default value of m = 3, under feasibility constraints provided by LIME. Features_to_vary were initially restricted to the instance's LIME-identified top-k features. DiCE-modified features were then considered in subsequent passes, again restricting them to the mutable-only feature set. Stability between LIME at x and its counterfactuals was assessed using cosine similarity, Spearman's rank correlation, and Jaccard overlap [31]–[34].

Table 2. LICE-guided model interventions

Intervention	Strategy	Action	Expected effect	Empirical outcome (Table 8)	Rationale
Reactive reweighting	Aggressive feature scaling (stress test)	Decrease over-resilient predictors and strengthen underutilized predictors to probe model sensitivity	Reduce FP by dampening over-resilient features; conduct a diagnostic stress test to assess robustness	Δ FN $+2095$ $\uparrow$, Δ FP -1148 $\downarrow$ diagnostic stress test only	LIME-derived importance used as direct guidance for influence reduction/increase; defines stability boundary.
Balanced reweighting	Mild feature scaling	Apply mild reduction to over-resilient features and strengthen underutilized under LICE guidance	Restore balance between FP and FN without destabilization	Accuracy $\approx$ 0.7545, AUC 0.8304 (stable) Δ FN -39, Δ FP -21;	Controlled perturbation validates safe incorporation of LICE-guided adjustments
Threshold calibration	Youden's J optimization	Choose $\tau = \arg\max(\mathrm{TPR} - \mathrm{FPR})$; optimal $\tau = 0.5696$	Improve sensitivity (FN$\downarrow$) without retraining	Δ FN -189, Δ FP $+179$; F1 $\approx$ 0.777	Explainability-driven threshold tuning adjusts the decision boundary to a clinically safer region
Integrated LICE optimization	Feature engineering + sample weighting + hyperparameter tuning + threshold adjustment	Add interpretable interaction; apply refined influence adjustments and Youden's J optimization	Integrate all prior learnings for stable recall gain and model reliability	Accuracy 0.7596, F1 0.7764, AUC 0.8309 (best trade-off) Δ FN -143, Δ FP $+119$;	Captures latent risk interactions and enhances sensitivity through interpretable feature increase

4. RESULTS AND DISCUSSION

In this section, the comparative performance of the proposed models is presented, and the improvements obtained through the LICE framework are discussed. Predictive strength, reliability, and explainability were evaluated using accuracy, AUC, recall, precision, F1-score, Cohen's Kappa, and MCC. The gradient boosting classifier reported in [24] is considered the baseline, against which the TunedGBDT represents the optimized benchmark.

The proposed LICE-EnhancedGBDT integrates local explanation through LIME with actionable counterfactuals through DiCE to refine misclassified cases, particularly FN. A set of conventional and ensemble classifiers, including LR, RF, NB, KNN, DT, SVM, and GBDT, was first benchmarked. Among these, GBDT achieved the strongest performance over the baseline.

The GBDT model was then further tuned and designated as TunedGBDT, serving as the benchmark, achieving an accuracy of 0.7558, an AUC of 0.8310, and an F1-score of 0.7708. Building on this benchmark, the LICE-guided model enhancement further improved the classifiers' recall to 0.8203 and F1 to 0.7764, while preserving the AUC at the same level, indicating that sensitivity increased without over-calibration. Table 3 shows the comparative results for all model variants.

Table 3. Performance comparison across predictive models

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
LR	0.7482	0.8242	0.7696	0.7476	0.7585	0.4961	0.4963
RF	0.7380	0.8066	0.7809	0.7283	0.7536	0.4747	0.4759
NB	0.7314	0.7901	0.7560	0.7302	0.7429	0.4619	0.4622
KNN	0.7380	0.7611	0.7256	0.7122	0.7186	0.4159	0.4160
DT	0.6396	0.6419	0.6296	0.6549	0.6420	0.2794	0.2796
SVM	0.7552	0.8258	0.8077	0.7394	0.7721	0.5088	0.5112
GBDT (baseline [24])	0.7549	0.8309	0.7949	0.7443	0.7686	0.5077	0.5090
Stacking	0.7549	0.6307	0.7874	0.7482	0.7673	0.3087	0.4094
TunedGBDT (benchmark)	0.7558	0.8310	0.8001	0.7436	0.7708	0.5120	0.5118
LICE-EnhancedGBDT	0.7596	0.8309	0.8203	0.7370	0.7764	0.5090	0.5204

Class-wise performance of LICE-EnhancedGBDT is presented in Table 4. Sensitivity (recall–for the positive diabetic class) = 0.8203 with precision (positive predictive value (PPV)) = 0.7370, resulting in F1 = 0.7764. Specificity (recall–for the negative (non-diabetic) class) = 0.6914 with precision (negative predictive value (NPV)) = 0.7849, resulting in F1-score = 0.7452, discrimination (AUC) = 0.8309, and accuracy = 0.7596, with a marginal increase in accuracy.

An operating point was selected that prioritizes sensitivity to minimize FN, i.e., missed diabetic cases, while maintaining PPV and NPV, which is an appropriate trade-off for clinical screening. Here, PPV and NPV depended on the cohort. The procedure followed to improve the model using XAI is explained in subsections 4.1. and 4.2.

Table 4. Class-wise performance of LICE-EnhancedGBDT

Class	Accuracy	AUC	F1	Recall	Precision
Positive (diabetic)	0.7596	0.8309	0.7764	0.8203 (sensitivity)	0.7370 (PPV)
Negative (non-diabetic)	0.7596	0.8309	0.7452	0.6914 (specificity)	0.7849 (NPV)

4.1. Systematic diagnosis of misclassifications

Misclassified cases in the benchmark model were systematically diagnosed using the LICE framework. A total of 1,956 FP and 1,417 FN were separated from the test set. Using LIME, local surrogate models were fitted around each misclassified instance to estimate feature contributions. These results were aggregated to obtain global influential features.

Table 5 presents the top five over-resilient and underutilized global influential features across FP and FN. Among the over-resilient features, $HeartDiseaseorAttack$ showed a strong negative contribution, indicating that it was likely to reduce a positive prediction. However, it was overridden by the contributions of other features.

Table 5. LIME's over-resilient and underutilized global influential features

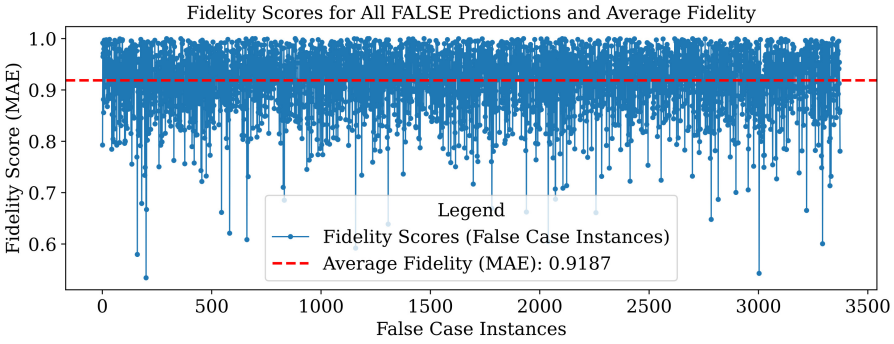
FP (over-resilient)		FN (underutilized)	
Feature	LIME weight	Feature	LIME Weight
$HighBP$	+0.0779	$HighBP$	-0.0306
$HighChol$	+0.0439	$HighChol$	-0.0232
$GenHlth$	+0.0317	$GenHlth$	-0.1447
Age	+0.0318	Age	-0.0318
$HeartDiseaseorAttack$	-0.0496	BMI	-0.0745

Figure 2 summarizes the fidelity-score distributions for all misclassified cases. Figures 2(a) to 2(c) show consistently high fidelity scores of 0.9187, 0.9193, and 0.9184, indicating that misclassifications were driven by the joint effect of over-resilient and underutilized features rather than poor surrogate fit. The analysis revealed that over-resilient features involving self-reported general health ($GenHlth$) and cardiovascular comorbidities ($HighBP$, $HeartDiseaseorAttack$), coupled with underutilized behavioral factors such as physical activity and fruit/vegetable intake, mirrored previously reported BRFSS behavioral risk trends [9], [27].

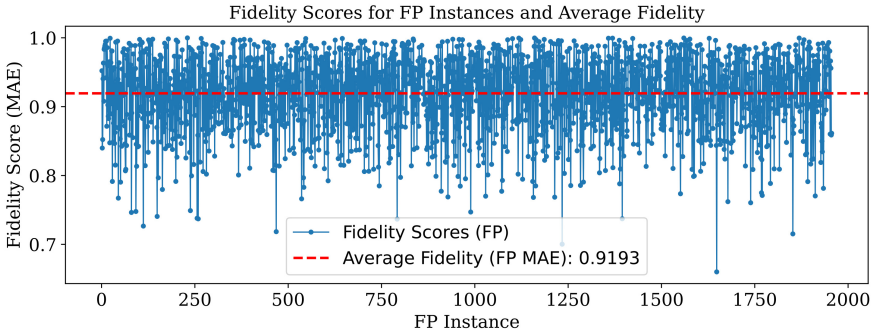
Counterfactuals were generated using DiCE, restricting perturbations to a "plausible feature set" (modifiable attributes such as $HighBP$, $HighChol$, $GenHlth$, BMI) while keeping immutable attributes (e.g., Age, $Income$) constant. Table 6 reports the stability metrics of these counterfactuals. High cosine similarity > 0.996 and Spearman rank correlation > 0.96 indicate stable local explanations, whereas moderate Jaccard overlap shows that LICE identified diverse yet actionable counterfactuals for misclassified cases.

4.2. Explanation-guided model improvement

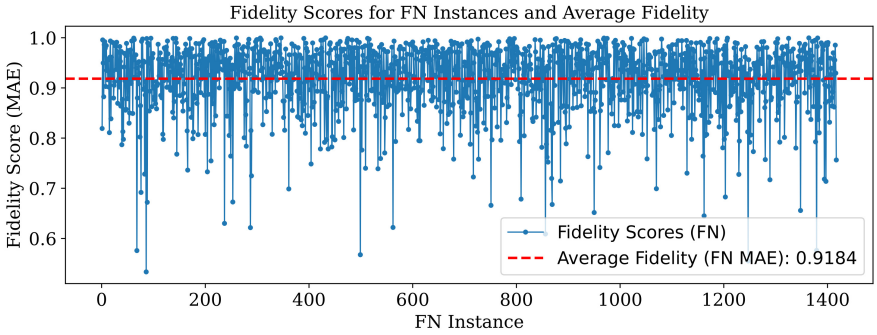
LICE's insights led to a sequence of four explanation-guided interventions. These interventions were progressively designed using reactive and balanced feature reweighting and threshold calibration, building an integrated LICE-Enhanced optimization. A comparison of the confusion matrices of TunedGBDT (benchmark) and LICE-EnhancedGBDT is shown in Table 7.



(a)



(b)



(c)

Figure 2. Fidelity scores and their average: (a) all false instances, (b) FP instances, and (c) FN instances

Table 6. Diverse counterfactual explanations' stability metrics across FP and FN instances

Metric	FP mean	FP ± SD	FN mean	FN ± SD
Cosine similarity	0.9966	±0.0036	0.9667	±0.0025
Spearman rank correlation	0.9713	±0.0353	0.9663	±0.0491
Jaccard overlap	0.3893	± 0.1371	0.2881	± 0.1361

Table 7. Confusion-matrix comparison and trade-off analysis

Model	TN	FP	FN	TP
TunedGBDT (benchmark)	4767	1956	1417	5672
LICE-enhancedGBDT	4648	2075	1274	5815
Δ (change)	-119	+119	-143	+143

The progress of each intervention is summarized in terms of both quantitative measures and qualitative interpretations in Table 8. Reactive reweighting: guided by the LICE systematic diagnosis of misclassification, reweighting was applied to model-sensitive features. From the FP cohort, LICE flagged $HighBP$ and $HighChol$ as over-resilient drivers; their influence was decreased by 20% each, and the cardiovascular risk-contributing feature, i.e., $HeartDiseaseorAttack$, was also addressed by increasing its influence by 10%. In the FN cohort, LICE found that $GenHlth$ and BMI were underutilized, and their influence was increased by 20% each. This intervention was marked as a stress test analysis. As expected, LICE-guided feature reweighting with high bidirectional changes resulted in model distortion and revealed a trade-off between sensitivity and specificity.

Table 8. Performance summary of explanation-guided interventions

Intervention	Strategy	Accuracy	F1	AUC	Δ FN	Δ FP	Interpretation
Reactive reweighting	Aggressive feature scaling	0.6872	0.6235	0.8200	+2095	-1148	Over-suppressed; diagnostic stress test only
Balanced reweighting	Mild feature scaling	0.7545	0.7686	0.8304	-39	-21	Stable under small perturbations; validated controlled use of LICE insights
Threshold calibration	Youden's J optimization	0.7565	0.7771	0.8300	-189	+179	Improved recall–precision balance via statistical tuning; no retraining
Integrated LICE optimization	Feature engineering + sample weighting + hyperparameter tuning + threshold adjustment	0.7596	0.7764	0.8309	-143	+119	Most balanced and interpretable configuration; 10% FN reduction

Balanced reweighting: building on the reactive stress test, small multipliers were applied to moderate LICE-identified effects while avoiding overcorrections. The influence of $HighBP$ and $HighChol$ for false alarms (FP) was decreased by 5%, the influence of $GenHlth$ and BMI for missed classifications (FN) was increased by 5%, and $HeartDiseaseorAttack$ was increased by 5% to emphasize cardiovascular risk. All other features were left unchanged. FN decreased by 39 instances (from 1,417 to 1,378) compared to the benchmark, and FP decreased by 21 cases (from 1,956 to 1,935). The model's overall performance remained stable under small perturbations. Similarly, accuracy = 0.7545 and F1 = 0.7686 were approximately equal to the benchmark. This controlled intervention demonstrated that LICE-guided feature reweighting with small bidirectional changes can simultaneously decrease both FP and FN while keeping the model's behavior unchanged at the operating point.

Threshold calibration: the benchmark model was trained with sample weights of 1.0 for class 0 and 1.5 for class 1 to reduce FN. Using out-of-fold training probabilities, a threshold of 0.5696 was selected by maximizing Youden's J (TPR-FPR) and applied unchanged to the test set. This yielded accuracy = 0.7565, F1 = 0.7771, and ROC-AUC = 0.8300; thresholding did not alter score rankings. FN decreased by 189 instances (from 1,417 to 1,228) compared to the benchmark, and FP increased by 179 cases (from 1,956 to 2,135). A shift towards sensitivity was observed, where the positive class recall increased to 0.83, accompanied by a modest increase in false alarms. Cost-sensitive training and threshold calibration effectively reduced the number of missed diabetic cases while maintaining strong overall model performance, despite a slight trade-off in specificity and PPV. Integrated LICE enhancement: here, two interpretable features were introduced: a LICE-flagged interaction, as in (1) and a cardiovascular-influenced feature, as in (2).

$$HighBP_HighChol = HighBP \times HighChol \tag{1}$$

$$HeartDiseaseorAttack_Weighted = HeartDiseaseorAttack \times 1.5 \tag{2}$$

With optimized hyperparameters ($learning_rate$ = 0.03, max_depth = 3, $n_estimators$ = 300, $subsample$ = 0.8), the proposed method maximized Youden's J on out-of-fold training predictions, selecting $\tau = 0.5696$ for the tuned model and $\tau = 0.5778$ after integration and retuning; each fixed threshold was then applied to the test set. The model achieved Accuracy = 0.7596, F1 = 0.7764, and AUC = 0.8309; and also reduced FN by 143 instances (from 1,417 to 1,274) compared to the benchmark, and FP increased by 119 instances (from 1,956 to 2,075).

5. CONCLUSION

This study worked beyond accuracy to improve the reliability and transparency of diabetes risk prediction by explanation-driven modeling. The proposed LICE framework successfully integrated local attribution (LIME) with feasibility-constrained counterfactual (DiCE) models. By stratifying errors, evaluating explanation quality, and translating diagnostics into interpretable interventions, LICE aligned model behavior with clinical priorities. LICE-EnhancedGBDT preserved global discrimination while raising sensitivity and F1, resulting in reduced FN. Cohort-level analysis also highlighted cardiovascular comorbidity indicators and modifiable behavioral factors, which informed our decision-making during implementation. Collectively, these findings showed that the explainability can be used not only to interpret but also to improve model behavior. The LICE framework is limited to a single cross-sectional dataset (BRFSS 2015). De-duplication alters the class balance slightly when compared to prior work. The LICE framework can be validated on additional datasets, such as Pima Indians diabetes database (PIMA), National Health and Nutrition Examination Survey (NHANES), and electronic health record (EHR), in addition to BRFSS, to quantify flexibility, calibration, and robustness under distribution shifts. This study can also extend the diagnostic intervention loop to other chronic conditions, such as hypertension and cardiovascular risk, and adapt the cohort-specific base to improve sensitivity and control false alarms.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Deepthi Angara	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Jhansi Yellapu	✓	✓		✓			✓			✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest. There are no known competing financial interests or personal relationships that could be perceived as influencing the work reported in this paper.

INFORMED CONSENT

This section is not applicable. This study involved secondary analysis of publicly available, fully anonymized BRFSS human survey data and did not involve direct interaction with human participants or the collection of personally identifiable information. All data used were publicly available and anonymized, and therefore no additional informed consent was required from individuals.

ETHICAL APPROVAL

This section is not applicable. This research did not involve direct recruitment of human subjects, human biological materials, or experimental procedures on animals. The work was conducted solely through secondary analysis of publicly available, anonymized BRFSS data and other non-sensitive data that did not require intervention with living organisms. Therefore, ethical approval from an institutional review board or

animal ethics committee was not necessary for this study.

DATA AVAILABILITY

We sourced our data from the publicly available behavioral risk factor surveillance system (BRFSS 2015), reference number [27]. The preprocessed data and analysis code supporting the findings of this work are not currently deposited in a public repository.

REFERENCES

[1] F. De-Giorgio, B. Benedetti, M. Mancino, E. Sala, and V. L. Pascali, "The need for balancing 'black box' systems and explainable artificial intelligence: a necessary implementation in radiology," *European Journal of Radiology*, vol. 185, 2025, doi: 10.1016/j.ejrad.2025.112014.

[2] F. F.-H. Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: applications, challenges, and AI-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, 2023, doi: 10.1080/15228053.2023.2233814.

[3] V. Bento, M. Kohler, P. Diaz, L. Mendoza, and M. A. Pacheco, "Improving deep learning performance by using explainable artificial intelligence (XAI) approaches," *Discover Artificial Intelligence*, vol. 1, no. 1, 2021, doi: 10.1007/s44163-021-00008-y.

[4] L. Longo *et al*., "Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, vol. 106, 2024, doi: 10.1016/j.inffus.2024.102301.

[5] S. Ali *et al*., "Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, 2023, doi: 10.1016/j.inffus.2023.101805.

[6] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021, doi: 10.1613/JAIR.1.12228.

[7] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable AI techniques in healthcare," *Sensors*, vol. 23, no. 2, Jan. 2023, doi: 10.3390/s23020634.

[8] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: a survey of needs, techniques, applications, and future direction," *Neurocomputing*, vol. 599, Sep. 2024, doi: 10.1016/j.neucom.2024.128111.

[9] World Health Organization, "Diabetes," who.int, 2024. Accessed: March 20, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[10] D. Majcherek, A. Ciesielski, and P. Sobczak, "AI-driven analysis of diabetes risk determinants in U.S. adults: exploring disease prevalence and health factors," *PLOS One*, vol. 20, no. 9, Sep. 2025, doi: 10.1371/journal.pone.0328655.

[11] U. Galicia-Garcia *et al*., "Pathophysiology of type 2 diabetes mellitus," *International Journal of Molecular Sciences*, vol. 21, no. 17, Aug. 2020, doi: 10.3390/ijms21176275.

[12] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13098-021-00767-9.

[13] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': explaining the predictions of any classifier," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, 2016, doi: 10.18653/v1/N16-3020.

[14] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617, doi: 10.1145/3351095.3372850.

[15] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, "Developing a fidelity evaluation approach for interpretable machine learning," Jun. 16, 2021, *arXiv:2106.08492*.

[16] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 2770–2824, 2024, doi: 10.1007/s10618-022-00831-6.

[17] I. Niño-Adan, D. Manjarres, I. Landa-Torres, and E. Portillo, "Feature weighting methods: a review," *Expert Systems with Applications*, vol. 184, 2021, doi: 10.1016/j.eswa.2021.115424.

[18] L. Cheng, K. R. Varshney, and H. Liu, "Socially responsible AI algorithms: issues, purposes, and challenges," *Journal of Artificial Intelligence Research*, vol. 71, pp. 1137–1181, 2021, doi: 10.1613/JAIR.1.12814.

[19] Y. Rong *et al*., "Towards human-centered explainable AI: a survey of user studies for model explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2104–2122, 2024, doi: 10.1109/TPAMI.2023.3331846.

[20] Z. Zhang, K. A. Ahmed, M. R. Hasan, T. Gedeon, and M. Z. Hossain, "A deep learning approach to diabetes diagnosis," in *Communications in Computer and Information Science*, 2024, pp. 87–99, doi: 10.1007/978-981-97-5937-8_8.

[21] S. Mansouri, S. Boulares, and S. Chabchoub, "Machine learning for early diabetes detection and diagnosis," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 15, no. 1, pp. 216–230, 2024, doi: 10.58346/JOWUA.2024.I1.015.

[22] A. F. Halias *et al*., "Type 2 diabetes mellitus prediction using data mining approach," in *2023 IEEE International Conference on Computing (ICOCO)*, Oct. 2023, pp. 29–34, doi: 10.1109/ICOCO59262.2023.10398078.

[23] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators," *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100118.

[24] R. Jose, F. Syed, A. Thomas, and M. Toma, "Cardiovascular health management in diabetic patients with machine-learning-driven predictions and interventions," *Applied Sciences*, vol. 14, no. 5, Mar. 2024, doi: 10.3390/app14052132.

[25] Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System (BRFSS)," cdc.gov. Accessed: March 20, 2025. [Online]. Available: https://www.cdc.gov/brfss/about/index.htm
[26] Centers for Disease Control and Prevention, "Survey data & documentation," cdc.gov. Accessed: March 20, 2025. [Online]. Available: https://www.cdc.gov/brfss/data_documentation/index.htm
[27] A. Teboul, "Diabetes health indicators dataset" Kaggle.com, 2021. Accessed: March 20, 2025. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
[28] C. O. Arimie, E. O. Biu, and M. A. Ijomah, "Outlier detection and effects on modeling," *Open Access Library Journal*, vol. 7, no. 9, pp. 1–30, 2020, doi: 10.4236/oalib.1106619.
[29] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
[30] M. D. Ruopp, N. J. Perkins, B. W. Whitcomb, and E. F. Schisterman, "Youden index and optimal cut-point estimated from observations affected by a lower limit of detection," *Biometrical Journal*, vol. 50, no. 3, pp. 419–430, Jun. 2008, doi: 10.1002/bimj.200710415.
[31] 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.
[32] A. Artelt *et al.*, "Evaluating robustness of counterfactual explanations," *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 – Proceedings*, 2021, pp. 1–9, doi: 10.1109/SSCI50451.2021.9660058.
[33] C. Spearman, "The proof and measurement of association between two things," *International Journal of Epidemiology*, vol. 39, no. 5, pp. 1137–1150, Oct. 2010, doi: 10.1093/ije/dyq191.
[34] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.

BIOGRAPHIES OF AUTHORS



Deepthi Angara is currently a research scholar in Computer Science and Engineering in the Department of Computer Science and Engineering at GITAM School of Technology (Deemed to be University), located in Visakhapatnam, Andhra Pradesh, India. She completed her M.Tech. in Computer Science and Systems Engineering at Andhra University in 2015. She is a highly motivated and proactive woman with a background in the tech industry and academia. She has a strong passion for working with data and possesses the ability to analyze and extract valuable insights from it through statistical analysis. With solid coding skills and experience in using Python for ML models, she is well-equipped to leverage data-driven approaches for prediction. Additionally, she possesses exceptional communication and interpersonal skills, which contribute to effective collaboration and knowledge sharing. She can be contacted at email: dangara@gitam.in.



Jhansi Yellapu earned her Ph.D. in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, India, in 2019. She completed her M.Tech. in Computer Science and Systems Engineering at Andhra University in 2008. Currently, she serves as an associate professor in the Department of Computer Science and Engineering at GITAM School of Technology (Deemed to be University), located in Visakhapatnam, Andhra Pradesh, India. With over 18 years of teaching experience and 12 years dedicated to research, her expertise spans machine learning, artificial intelligence, data engineering, and image processing. She is deeply engaged in numerous research initiatives and has a strong publication record in esteemed journals and conferences. Her dedication to pushing the boundaries of knowledge in her field is well recognized. Additionally, she is highly committed to fostering her students' academic and professional development, making her a respected figure both in research and education. She can be contacted at email: jyellapu@gitam.edu.