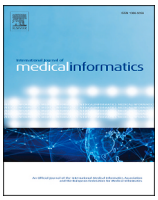
International Journal of Medical Informatics 216 (2026) 106488

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Original research article

Explainable TabNet for gestational diabetes prediction with physician-in-the-loop and multi-site clinical validation

Anju Narayanan [a,*], Praveen Sankaran [a], Uma V. Sankar [b], Simi Kurian [c], Teslin John [d]

[a] Department of Electronics and Communication Engineering, NIT Calicut, Kerala, 673601, India
[b] Aster Medcity, Ernakulam, Kerala, India
[c] Pariyaram Medical College, Kannur, Kerala, India
[d] Sacred Heart Hospital, Palakkad, Kerala, India

HIGHLIGHTS

- TabNet achieves 97.13% accuracy for GDM prediction using routine clinical variables.
- Three-stage hybrid imputation improves F1-score by 4.98% over mean imputation.
- Dual-stage physician validation yields 96.7% concordance and kappa of 0.909.
- Prospective multi-site validation across three Kerala hospitals confirms generalizability.
- SHAP reveals PCOS-prediabetes synergy as dominant combined GDM risk signal.

ARTICLE INFO

Keywords:
Gestational diabetes mellitus
Deep learning
Explainable AI
TabNet
Physician-in-the-loop
Clinical validation
Multi-site validation

ABSTRACT

Background: Gestational diabetes mellitus (GDM) affects 15–25% of pregnancies worldwide and poses serious risks of macrosomia, preeclampsia, neonatal hypoglycaemia, and long-term type 2 diabetes. Existing machine learning models lack prospective multi-site external validation and formal physician trust evaluation, limiting real-world applicability.

Objectives: To develop a clinically validated, explainable deep learning framework for GDM prediction using routinely available first-antenatal-visit clinical features, and to evaluate clinical readiness through dual-stage physician-in-the-loop (PITL) validation.

Methods: A TabNet binary classifier was developed on 3,525 clinical records using a three-stage feature-tailored hybrid imputation strategy (GAIN for HDL and OGTT; MissForest for Systolic BP; Mean for BMI). To prevent data leakage, SMOTE-based class balancing was applied exclusively within the training folds of a 5-fold stratified cross-validation pipeline, with validation folds remaining untouched. Explainability was delivered through TabNet intrinsic feature masks, SHAP, and LIME. Two-stage clinical validation comprised: (1) blinded PITL review by four certified obstetricians evaluating 30 patient cases with XAI explanations; and (2) prospective external validation across three independent Kerala hospitals totaling 80 patients.

Results: The proposed TabNet model achieved 97.13% accuracy, 94.05% precision, 98.91% recall, and 96.22% F1-score, outperforming ten baseline classifiers including Random Forest, XGBoost, and SVM under identical preprocessing conditions. Compared to recent state-of-the-art GDM prediction studies, the proposed model consistently outperformed comparable methods—under a rigorous 5-fold cross-validation strategy with confidence intervals, while most existing studies rely on single train–test splits without cross-validation. PITL validation yielded 96.7% concordance, an average Cohen's kappa of 0.909, and Fleiss' kappa of 0.963, with no prior GDM study reporting such formal physician endorsement. External multi-site F1 scores ranged from 83.70% to 87.00% across all three hospitals, reflecting an expected performance reduction in prospective real-world data, partly attributed to inter-site variability in feature availability and clinical data recording protocols. SHAP analysis identified a strong model-level interaction between PCOS and prediabetes as the dominant combined GDM risk signals, independently corroborated by all four obstetricians.

* Corresponding author.
Email address: anjunarayanan41@gmail.com (A. Narayanan).

https://doi.org/10.1016/j.ijmedinf.2026.106488
Received 3 March 2026; Received in revised form 2 May 2026; Accepted 13 May 2026
Available online 16 May 2026

Conclusion: The proposed framework integrates explainable deep learning with prospective dual-stage clinical validation, demonstrating promising performance as a clinically oriented proof-of-concept for the assessment of risk of GDM using routine clinical variables.

Trial Registration: Clinical Trials Registry India, CTRI/2024/08/073158.

1. Introduction

Gestational diabetes mellitus (GDM) affects 15–25% of pregnancies worldwide [1] and arises primarily from hormonal changes during pregnancy. Placental hormones including human placental lactogen, estrogen, progesterone, and cortisol increase significantly in the second and third trimesters, inducing insulin resistance to ensure adequate fetal glucose supply. When the pancreas fails to compensate with sufficient insulin production, maternal hyperglycemia results [2]. Left unmanaged, GDM leads to serious short- and long-term complications for both mother and child, including macrosomia, preeclampsia, neonatal hypoglycemia, increased cesarean delivery rates, and elevated risk of type 2 diabetes [3]. Early diagnosis, personalized interventions, and lifestyle modifications are essential to mitigating these risks.

GDM is typically diagnosed via the Oral Glucose Tolerance Test (OGTT) [4] at 24–28 weeks, the current gold standard; however, its requirement for fasting and multiple blood draws makes it time-consuming and uncomfortable, frequently delaying timely intervention. Rule-based risk factor models offer an alternative but are limited by moderate accuracy, poor generalizability, and late detection. Machine learning (ML) addresses these gaps by learning complex patterns from clinical, biochemical, and lifestyle data, with commonly employed approaches including random forest [5], support vector machines [6], gradient boosting methods such as XGBoost and LightGBM [7,8], and neural networks [9].

The quality and representativeness of the dataset are fundamental to ML model performance. Since ML algorithms are inherently data-driven, model accuracy, robustness, and generalizability depend critically on data completeness, consistency, and representativeness. Issues such as missing values, incorrect measurements, class imbalance, and outliers can introduce bias and reduce clinical applicability. These challenges are typically addressed through preprocessing steps including missing value imputation, class rebalancing, outlier detection, and normalization, though the effectiveness of each strategy depends on the distributional characteristics of the specific dataset.

GDM datasets combine numeric, categorical, and binary variables [10], with missing values arising from inconsistent recording, entry errors, and varying clinical protocols, particularly critical for key biomarkers such as OGTT, PCOS, and BMI. Standard imputation approaches span a capability hierarchy: simple methods (mean [11], median [12], mode [13]) are computationally efficient but ignore inter-feature relationships; KNN and regression-based methods capture linear correlations but introduce bias under nonlinearity; MICE [14] handles uncertainty through iterative conditional estimation; and tree-based (MissForest [15,16]) and deep learning (GAIN [17]) approaches progressively capture nonlinear interactions and complex biochemical dependencies. Since GDM features exhibit heterogeneous distributional properties, no single imputation method suffices universally; the selection of an appropriate strategy must therefore be guided by each feature's missingness pattern and distributional characteristics, rather than applying a uniform approach.

Class imbalance presents another key challenge, as GDM-positive cases are typically underrepresented [18]. Oversampling techniques including SMOTE [19], random oversampling [20], and ADASYN [21] have been explored in prior work, each offering different trade-offs between synthetic sample realism and minority class coverage.

Advanced deep learning architectures have shown strong performance across diverse biomedical data handling problems. For instance, capsule network-based models such as iSucc-SnCNs [22], DeepAIPs-SFLA [23], and pNPs-CapsNet [24] have been successfully applied to sequence-based biological classification tasks such as succinylation site prediction, anti-inflammatory peptide identification, and neuropeptide prediction, respectively. While these architectures are specifically designed for sequence-based biological inputs and are not directly applicable to structured clinical tabular data for GDM prediction, they reflect the broader trend of leveraging deep learning for complex biomedical classification problems.

For structured clinical tabular data—the data type encountered in GDM prediction—TabNet [25] represents a particularly suitable deep learning architecture. Unlike conventional deep learning models that treat all features equally and require post-hoc explanation tools, TabNet employs a sequential attention mechanism that performs instance-wise sparse feature selection at each decision step, making it inherently interpretable for tabular clinical data [26]. This built-in interpretability is a critical advantage for clinical decision support, as it enables direct understanding of which features drove each individual prediction without relying solely on external explanation tools. To further enhance transparency, three complementary explainability methods are employed as key outcomes: TabNet's intrinsic feature masks for global and local feature attribution, SHAP [27] for unified feature contribution analysis and interaction detection, and LIME [28] for patient-level local explanations.

Several recent studies have applied XAI to GDM prediction using SHAP or LIME as post-hoc tools alongside conventional classifiers [29–31]. However, these approaches share notable limitations: reliance on single train–test splits without rigorous cross-validation, absence of prospective external validation across independent clinical sites, and lack of formal physician evaluation of XAI outputs. Critically, no prior study has combined intrinsic model explainability with dual-stage clinical validation involving blinded physician review and prospective multi-site testing.

To address these gaps, this study proposes a TabNet-based [25] explainable deep learning framework for GDM risk assessment using routinely collected first-antenatal-visit clinical variables. It is important to clarify that while OGTT is included as one of the input features when available at the first antenatal visit, the framework is designed to operate on the broader set of routinely collected clinical and lifestyle variables, supporting early risk stratification rather than replacing formal OGTT-based diagnosis at 24–28 weeks. While the training dataset originates from a single public source and external validation is geographically constrained to Kerala, India, this work represents a rigorous proof-of-concept establishing the feasibility of clinically trustworthy, interpretable GDM screening using routine variables. Broader multi-ethnic validation remains an essential direction for future work.

The proposed framework advances beyond existing explainable GDM models in four specific ways that have not been collectively reported in prior work: (1) a **feature-tailored hybrid imputation strategy** is designed based on each variable's missingness pattern and biochemical characteristics, rather than applying a uniform method; (2) **leakage-free SMOTE** is applied strictly within training folds to prevent data contamination; (3) a **tri-method XAI framework** combining TabNet masks, SHAP, and LIME provides both global and patient-level interpretability; and (4) a **dual-stage PITL validation** comprising blinded expert review by four certified obstetricians and prospective external validation across three geographically independent hospitals—a validation design not previously reported for GDM prediction models.

The specific contributions of this research are as follows:

- **Feature-tailored hybrid imputation:** A three-stage strategy combining GAIN (for HDL and OGTT), MissForest (for Systolic BP), and Mean imputation (for BMI) is developed, with each method selected based on the missingness pattern and distributional properties of the respective feature, rather than applying a single uniform imputation approach.
- **Explainable TabNet classifier:** A TabNet binary classifier is developed and validated for GDM prediction, leveraging intrinsic sparse feature masks for instance-wise interpretability at each decision step, without sole reliance on post-hoc explanation tools.
- **Tri-method XAI framework:** TabNet feature masks, SHAP, and LIME are systematically integrated to provide consistent global and patient-level feature attribution, enabling clinically interpretable and cross-validated risk explanations—an outcome of the GDM prediction framework rather than its primary objective.
- **Dual-stage physician-in-the-loop (PITL) validation:** A formal PITL protocol is implemented comprising blinded independent review by four certified obstetricians and prospective external validation across three independent hospitals in Kerala, India—a dual-stage clinical validation design not previously reported for GDM prediction models.
- **Rigorous leakage-free evaluation pipeline:** SMOTE oversampling is applied exclusively within training folds of a 5-fold stratified cross-validation framework, ensuring unbiased performance estimation and preventing contamination of validation sets by synthetic samples.

2. Proposed method

This work proposes a hybrid imputation technique that includes mean imputation, MissForest, and GAIN, taking into account the underlying missing pattern types in the GDM dataset [32]. To reduce class disparity, SMOTE is used. Following this, a TabNet deep learning network is used to predict the risk of GDM in a patient. In addition to the inherent interpretability in TabNet, external explainability techniques such as SHAP and LIME are applied to provide global and local feature-level explanations. Finally, external validation is conducted at three independent hospitals, and a physicians-in-the-loop validation approach is used to evaluate the clinical relevance of the outcomes.

2.1. Dataset and dataset preprocessing

2.1.1. Dataset

A publicly available clinical dataset [32] from Anbu Hospital and Nalam Clinic, Kumbakonam, Tamil Nadu, curated in 2021 is used for model development. This dataset contains 3,525 patient records with 16 predictive features and a binary GDM class label (38.92% GDM, 61.08% non-GDM). The dataset includes 9 numerical features (age, BMI, HDL, Sys_BP, Dia_BP, OGTT, haemoglobin, gestation in previous pregnancy) and 7 categorical features (family history, sedentary lifestyle, PCOS, prediabetes, unexplained prenatal loss, large child or birth defect). Feature descriptions and variable summaries with missing data rates are provided in Supplementary Tables S1 and S2.

It is important to note that this publicly available retrospective dataset was used exclusively for model training and internal validation. The prospective external validation was conducted separately across three independent hospitals in Kerala, India (Pariyaram Medical College, SH Hospital, and Aster Medcity), involving 80 newly recruited participants under formal ethical approval and informed consent, as described in Section 2.1.2. These two data sources are entirely distinct and were never combined during any stage of model development or evaluation.

2.1.2. Ethical approval and trial registration

This study is part of a registered clinical trial (CTRI/2024/08/073158). Ethical approval was obtained from the Institutional Ethics Committee, Aster Medcity, Kochi (Ref: AM/EC/367-2023, dated 23.12.2023). All procedures conformed to the Declaration of Helsinki. Informed consent was obtained from all prospective external validation participants, and all patient data were fully anonymised prior to analysis.

The training dataset comprises publicly available, fully de-identified retrospective records requiring no additional consent. The prospective external validation data were collected under formal ethics approval from the respective governing bodies of all three participating hospitals. This clear separation between retrospective training data and prospectively collected validation data ensures the integrity of the clinical evaluation.

2.1.3. Dataset preprocessing

The class label (GDM/Non-GDM) is encoded for model compatibility, with GDM coded as 1 and non-GDM as 0. Missing values were first addressed using simple mean imputation as a baseline, yielding a TabNet accuracy of 93.12%, precision of 90.12%, recall of 92.76%, and F1-score of 91.24%. To effectively address missing values and minimize information loss, a three-stage feature-tailored hybrid imputation strategy was subsequently developed. The selection criteria for each imputation method were based on two factors: (1) the missingness rate of the feature, and (2) the distributional and biochemical characteristics of the variable. Specifically:

- **GAIN** [33] was selected for **HDL** (28.4% missing) and **OGTT** (14.6% missing), as these biomarkers exhibit complex nonlinear biochemical interdependencies that require deep learning-based imputation to capture adequately.
- **MissForest** [16] was selected for **Systolic BP** (48.4% missing), as its very high missingness rate combined with nonlinear inter-feature correlations makes it unsuitable for simpler imputation approaches. MissForest's iterative random forest mechanism is robust to high missingness and mixed-type data.
- **Mean imputation** [34] was selected for **BMI** (30.7% missing), as BMI exhibits a relatively symmetric distribution in the dataset, making mean-based imputation a clinically reasonable and computationally efficient choice.

Each method is detailed in Supplementary Section S2. The proposed hybrid strategy substantially improved all performance metrics, achieving 97.13% accuracy, 94.05% precision, 98.91% recall, and 96.22% F1-score—corresponding to gains of 4.01%, 3.93%, 6.15%, and 4.98% respectively over simple imputation, confirming the clinical importance of feature-tailored missing value handling. A detailed comparison of TabNet performance across all ten imputation strategies is provided in Supplementary Table S3. Fig. 1 illustrates the feature-wise missing rates and selected imputation strategies.

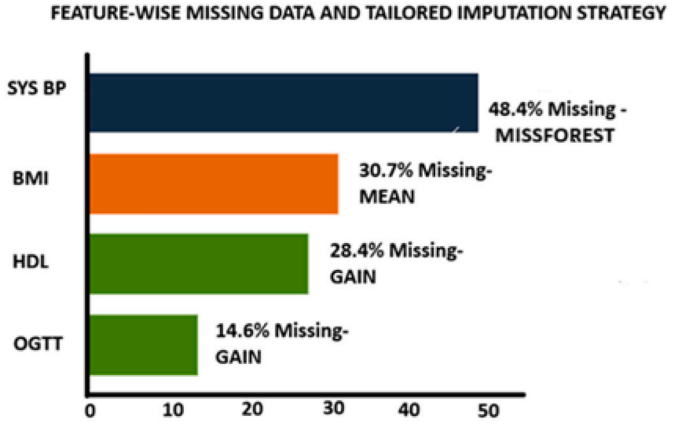


Fig. 1. Feature-wise missing data rates and tailored imputation strategies. Methods are selected based on missingness profile and feature characteristics.

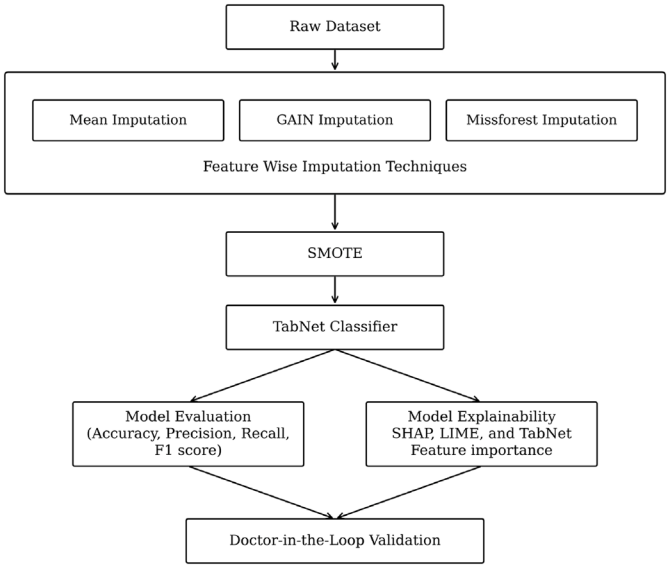


Fig. 2. Workflow summarizing the complete analytical pipeline: feature-tailored hybrid imputation (GAIN, MissForest, Mean), stratified 5-fold cross-validation, SMOTE resampling within training folds only, TabNet classification, and dual-stage doctor-in-the-loop validation.

Cross-validation and oversampling pipeline. To ensure full analytical transparency, the complete preprocessing pipeline follows a strict sequential order to prevent data leakage: (1) feature-tailored imputation is applied to the full dataset; (2) the imputed dataset is split into 5 stratified folds; (3) SMOTE oversampling is applied exclusively within each training fold; and (4) the TabNet model is trained on the resampled training fold and evaluated on the untouched validation fold. This ensures that no synthetic samples or imputation statistics from the training set contaminate the validation set at any stage.

In each fold, four subsets are used for training and one for validation, ensuring that each sample is validated exactly once. Stratification preserves the original class distribution of GDM and non-GDM cases across folds, resulting in robust, generalizable performance estimates.

Multiple oversampling techniques, including SMOTE (Synthetic Minority Oversampling Technique) [35], ADASYN (Adaptive Synthetic Sampling) [36], and random oversampling [37], are systematically evaluated. Experimental results show that SMOTE delivered superior results with improved sensitivity towards the minority class. Therefore, SMOTE is used within each training fold of the cross-validation pipeline, with the validation set remaining untouched to preserve evaluation integrity [19]. While cost-sensitive learning represents a complementary alternative to oversampling that avoids synthetic data generation entirely, its comparison with SMOTE-based approaches is noted as a direction for future work. A workflow summarizing hybrid imputation, SMOTE resampling, and the modeling pipeline for GDM prediction and doctor-in-the-loop validation is given in Fig. 2.

2.2. Data classification

TabNet, a deep learning architecture for tabular data problems, is utilized to develop the binary prediction model. TabNet uses a sequential attention mechanism for selecting the important features for each individual instance, rather than depending on a fixed global feature importance across all data points [25]. Unlike conventional deep learning models that require post-hoc explanation tools, TabNet provides built-in interpretability through sparse feature masks generated at each decision step, making it particularly suitable for clinical decision support tasks such as GDM prediction [26] (Fig. 3).

The architecture has multiple sequential steps where the outputs of each step are given to the next step as input. For each step, three key components are involved: the feature transformer, the attentive transformer, and the mask. This design enables TabNet to perform stepwise instance-specific feature selection and representation learning. To prevent overfitting, the proposed pipeline incorporates three complementary mechanisms: (1) a 5-fold stratified cross-validation framework

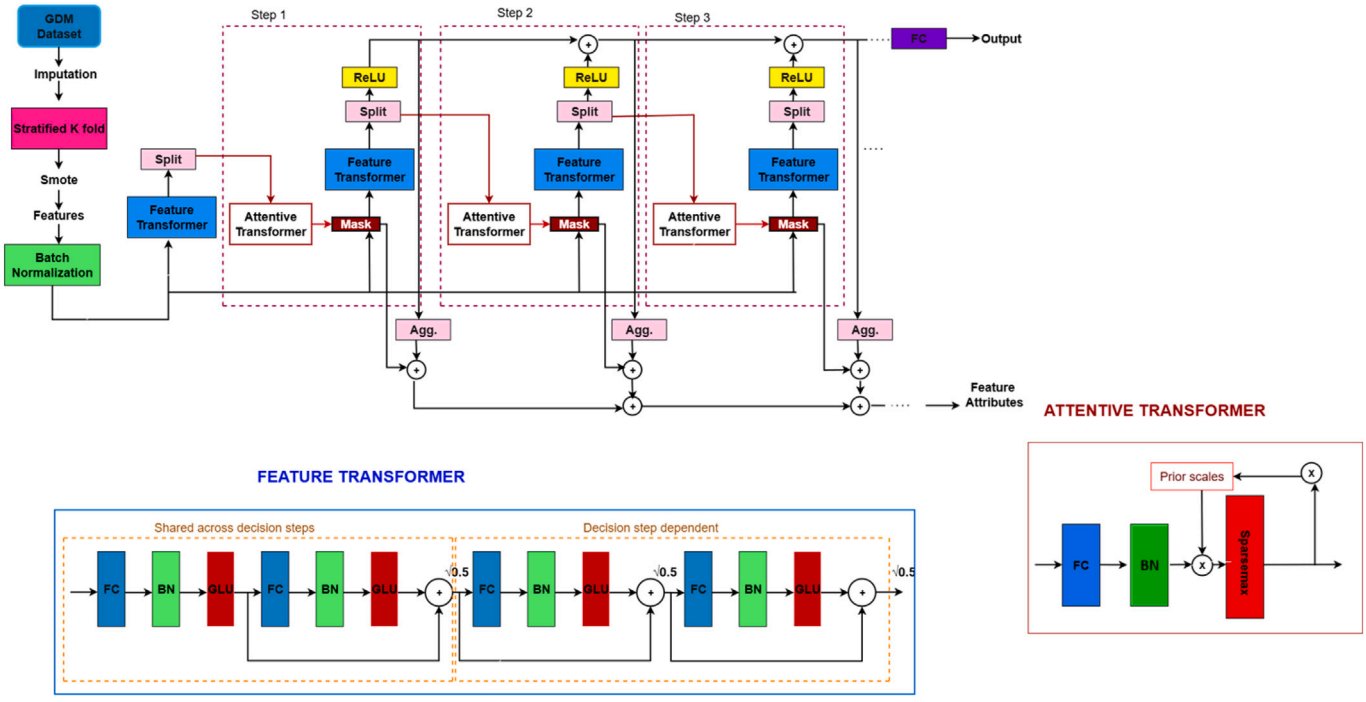


Fig. 3. Proposed architecture: TabNet Encoder with SMOTE for GDM prediction. The pipeline includes dataset preprocessing, stratified k-fold splitting, SMOTE balancing, and a TabNet encoder with three decision steps. Each step applies sparse feature selection, transformation, and aggregation, enabling interpretable and accurate classification.

Table 1
TabNet hyperparameter configuration determined via Optuna Bayesian optimization.

Hyperparameter	Value	Role
N_d (Embedding dim.)	64	Width of decision step output
N_a (Attention dim.)	64	Width of attention embedding
N_steps	5	Number of sequential decision steps
Gamma (γ)	1.3	Feature reuse relaxation factor
Lambda sparse	0.001	Sparsity regularization coefficient
Learning rate	0.02	Optimizer step size
Batch size	1024	Training batch size
Optimizer	Adam	Gradient descent optimizer
Max epochs	200	Training iterations
Patience	15	Early stopping patience

ensuring performance is evaluated on held-out data never seen during training; (2) SMOTE oversampling applied exclusively within training folds, preventing synthetic samples from contaminating validation sets; and (3) a sparsity regularization loss term (L_{sparse}) embedded within the TabNet architecture that penalizes the model for attending to excessive features simultaneously, acting as an implicit regularizer. The consistent performance across all five folds confirms that the model generalizes well without overfitting. The TabNet classifier was configured with optimized hyperparameters determined via Optuna Bayesian optimization [25]. The final configuration is summarized in Table 1.

2.2.1. Feature transformer

The *Feature Transformer* is the core component of TabNet. For GDM prediction, it takes all raw input features, such as BMI, OGTT, and HDL, and converts them into higher-level representations for decision-making. This block consists of a stack of fully connected (FC) layers interleaved with Batch Normalization (BN), Gated Linear Units (GLUs), and residual (skip) connections.

GLUs act as dynamic filters in the network. They multiply the main input by a sigmoid gate, allowing only informative signals (here, variables most predictive of GDM status) to pass while suppressing the irrelevant ones. Batch normalization is used to stabilise learning by maintaining a consistent input distribution throughout training. Residual connections further enhance network stability by mitigating the vanishing gradient problem, enabling direct gradient flow from later layers to earlier ones.

2.2.2. Attentive transformer

The *Attentive Transformer* is responsible for instance-wise feature selection at each decision step. It operates based on two key mechanisms:

- **Sparsemax Activation:** Unlike softmax, which always generates a dense probability distribution where all classes are assigned some probability, sparsemax generates sparse probability vectors by assigning exactly zero probability to less relevant classes.
- **Learnable Feature Mask:** At each step, TabNet generates a learnable sparse feature mask from the current input activation. This mask is unique and determines the most relevant features for processing at that step.

The following equations represent the core computations in the TabNet.

Input Feature Masking,

$$x_t^M = x \odot M_t \tag{1}$$

where x is the raw input feature vector, M_t is the feature mask at the decision step t, and $\odot$ denotes element-wise multiplication.

Feature Transformer (Step 0)

$$z_0 = \text{BN}(x_t^M W_0) \tag{2}$$

where W_0 is the weight matrix of the initial linear transformation, and BN denotes batch normalization.

Gated Linear Unit (GLU) for Layer l

$$z_l = \left(\text{BN}(z_{l-1} W_l^a)\right) \odot \sigma\left(\text{BN}(z_{l-1} W_l^b)\right) \tag{3}$$

where W_l^a and W_l^b are learnable weight matrices for the linear and gating paths, and $\sigma(\cdot)$ is the sigmoid function.

Residual Connection

$$z_l = \alpha z_l + (1-\alpha) z_{l-1}, \quad \alpha \in [0,1] \tag{4}$$

In our model, $\alpha = 1$, to fully rely only on the output of the current layer without residual mixing.

$$z_l = z_l + z_{l-1} \tag{5}$$

Attentive Transformer is responsible for selecting the most relevant features at each decision step. Masking scores are obtained using a fully connected layer,

$$H = \text{BN}(W_a \cdot X + b_a) \tag{6}$$

Sparsemax or relaxed softmax is applied to get the feature mask:

$$M = \text{sparsemax}(H \odot P_{t-1}) \tag{7}$$

The prior scale P is updated as,

$$P_t = P_{t-1} \odot (\gamma - M_t) \tag{8}$$

where P_t is the prior feature importance at step t, $\gamma > 1$ is the relaxation parameter to reduce feature reuse, and M_t is the current mask. In this work, $\gamma = 1.3$. The final prediction aggregation is obtained as,

$$y = \sigma\left(\sum_{t=1}^{T} d_t\right) \tag{9}$$

where d_t is the output of the decision step t, and $\sigma(\cdot)$ is the sigmoid activation function used for binary classification. A loss term L_{sparse} is added during training to penalize the model when it uses too many features,

$$L_{\text{sparse}} = \lambda_{\text{sparse}} \sum_{t=1}^{T} \text{mean}\left(\sum M_t \odot \ln(M_t + \varepsilon)\right) \tag{10}$$

where λ_{sparse} is the regularization coefficient controlling the sparsity strength, T is the total number of decision steps in the TabNet architecture, M_t is the feature selection mask at decision step t, and ε is the small positive constant added for numerical stability in log.

$$\text{BCE} = -\left[y \cdot \ln(\hat{y}) + (1-y) \cdot \ln(1-\hat{y})\right] \tag{11}$$

where BCE denotes the binary cross-entropy loss, y is the true binary label (0 or 1), $\hat{y}$ is the predicted probability of the positive class (i.e., model output).

$$L_{\text{total}} = L_{\text{BCE}} + L_{\text{sparse}} \tag{12}$$

where L_{total} denotes the final loss used for optimization, L_{BCE} represents the Binary Cross-Entropy, and L_{sparse} is the sparsity-inducing regularization loss that encourages minimal feature selection.

For the proposed GDM prediction task, the TabNet Classifier was configured with optimized hyperparameters determined via Optuna Bayesian optimization. The final configuration (embedding/attention dimensions = 64, decision steps = 5, sparsity penalty = 0.001, learning rate = 0.02) was selected based on validation performance. Complete hyperparameter details and their roles are provided in Supplementary Table S4.

2.3. Explainable artificial intelligence (XAI) framework

To ensure clinical transparency and trust, the proposed model integrates three complementary explainability techniques: TabNet's intrinsic feature masks, SHAP (SHapley Additive exPlanations) [27], and LIME (Local Interpretable Model-Agnostic Explanations) [28]. Each method provides distinct yet complementary insights into model behavior at both global (population) and local (patient) levels. A structured comparison of all three XAI methods covering interpretation type and clinical utility is provided in Supplementary Table S5.

A) Intrinsic explainability via TabNet feature masks

TabNet provides built-in interpretability through sparse feature masks generated at each decision step. These instance-wise masks identify the most influential features contributing to each prediction without requiring post-hoc tools. By aggregating these masks across all patients, population-level risk factors for GDM are identified. This mechanism enables both patient-specific explanations (local interpretability) and population-level insights (global interpretability). The sequential decision process allows the model to focus on commonly relevant features in early steps while progressively shifting attention to more specific features in later steps.

B) SHAP for feature attribution

SHAP provides unified feature attribution by computing each feature's contribution to individual predictions based on cooperative game theory. Global interpretability is achieved through beeswarm summary plots and feature importance rankings, while local explanations are visualized using waterfall plots. Additionally, SHAP dependency plots reveal nonlinear relationships and feature interactions among top predictors, offering deeper clinical insights into risk factor synergies. Detailed mathematical formulations of SHAP are provided in Supplementary Section S5.

C) LIME for local explanations

LIME generates local explanations by fitting interpretable surrogate models around individual predictions. For each patient instance, LIME creates synthetic perturbations of the input, weights them by proximity to the original instance, and trains a sparse linear model that approximates the TabNet behavior locally. This provides case-specific feature contributions that are easily interpretable by clinicians. The complete LIME methodology, including its architectural framework, is detailed in Supplementary Section S5.

3. Results

3.1. Evaluation metrics

Standard classification metrics (accuracy, precision, recall, F1-score) are used, with recall and F1-score being especially critical for SMOTE evaluation [38]. Metric definitions are provided in Supplementary Tables S6 and S7. Statistical agreement metrics (percentage concordance, Cohen's kappa [39,40], Fleiss' kappa [41]) used for PITL validation are summarized in Supplementary Table S8. All performance metrics are reported with 95% confidence intervals derived from 5-fold cross-validation results to support statistical interpretation of model comparisons.

3.2. GDM prediction performance

SMOTE achieved the best performance among three oversampling methods evaluated (Table 2). The full pipeline (GAIN + MissForest + Mean imputation + SMOTE + TabNet) achieved 97.13% accuracy, 94.05% precision, 98.91% recall, and 96.22% F1-score.

Model calibration was assessed using the Brier score. The proposed model achieved a Brier score of 0.0265 (95% CI: 0.018–0.035), indicating excellent probability calibration and confirming that the model's confidence estimates are reliable for clinical decision support.

Table 2
Performance Comparison of oversampling techniques with TabNet.

Method	Acc (%)	Prec(%)	Rec (%)	F1 (%)
SMOTE	97.13	94.05	98.91	96.22
ADASYN	95.83	93.22	95.47	94.33
Random Oversampling	94.65	92.78	94.02	93.38

Table 3
Performance Comparison of Models (mean ± 95% CI across 5-fold cross-validation).

Rank	Model	Acc	Prec	Rec	F1
1	TabNet	97.13±0.41	94.05±0.53	98.91±0.61	96.22±0.50
2	Random Forest	95.54±0.48	94.03±0.51	97.17±0.55	95.57±0.49
3	SVM	95.43±0.50	93.81±0.54	97.17±0.58	95.46±0.52
4	XGBoost	95.43±0.47	94.23±0.52	96.70±0.56	95.45±0.50
5	CatBoost	95.20±0.49	93.18±0.55	97.40±0.57	95.24±0.51
6	LightGBM	95.20±0.51	94.01±0.53	96.47±0.59	95.22±0.53
7	KNN	95.08±0.52	93.58±0.56	96.70±0.60	95.12±0.54
8	FNN	95.08±0.50	93.79±0.54	96.47±0.58	95.11±0.52
9	AdaBoost	94.73±0.53	93.76±0.57	95.78±0.61	94.76±0.55
10	Logistic Regression	94.62±0.55	93.55±0.58	95.78±0.62	94.65±0.57
11	Decision Tree	94.50±0.57	93.54±0.60	95.80±0.63	94.53±0.58

Calibration plots (reliability diagrams) are provided in Supplementary Figure S6 for a more comprehensive visual assessment of predicted probability alignment with observed outcomes.

3.3. Performance comparison with baseline models

Table 3 presents a performance comparison of the proposed TabNet model with various machine learning models, all trained using the same preprocessing pipeline (hybrid imputation + SMOTE). The proposed TabNet model achieved the highest accuracy (97.13% ± 0.41%), recall (98.91% ± 0.61%), and F1-score (96.22% ± 0.50%) among all models. Confidence intervals were derived from 5-fold cross-validation results across all models to ensure statistical comparability.

Among the compared models, Random Forest, SVM, XGBoost, and CatBoost also performed well, but their F1-scores were consistently lower than TabNet across all folds. Traditional models such as logistic regression and decision tree showed comparatively lower performance, suggesting reduced capacity to capture complex nonlinear patterns in the GDM dataset. Deep learning models such as the feed-forward neural network gave competitive results but were less effective than TabNet on this tabular data. Overall, TabNet offers a consistent combination of accuracy and robustness across all cross-validation folds.

3.4. Comparison with state-of-the-art methods

Table 4 presents a comparative analysis of the proposed TabNet framework with recent state-of-the-art GDM prediction studies. Several research groups report strong performance, particularly Hassan et al. [29] (Acc 98.21%) and Ji et al. [42] (AUC 0.984). However, many of these studies rely on single train–test splits, small sample sizes, costly biomarkers, or single-centre retrospective data, which limit real-world generalizability. Large population-based studies such as Watanabe et al. [43] and Belsti et al. [44] demonstrate improved robustness but achieve comparatively lower predictive performance. In contrast, the proposed model reports performance with 95% confidence intervals across a rigorous 5-fold cross-validation strategy, providing a statistically reliable estimate of clinical performance. It is acknowledged that the gap between internal validation (F1: 96.22%) and external validation (F1: 83.70–87.00%) reflects the expected performance reduction when applying a model trained on a single-source dataset to prospective real-world clinical data from independent sites, and indicates the need for broader multi-site training in future work.

Table 4
Comparison of machine learning-based GDM prediction studies.

Study	Dataset / Population	N	Features	Best Model (XAI)	Split	Performance	Limitations
Ji et al. [42]	Hainan Provincial People's Hospital, China	89	11 metabolites + 3 clinical	MLP (SHAP)	70/30	AUC 0.984; Acc 92.6%; Prec 93.8%; Rec 92.3%	Very small, single-center; no external validation; high metabolomics cost
Abe et al. [45]	State Govt. Hospitals, Nigeria	5981	Demographics, pregnancy, FPG, symptoms	Voting Ensemble (No XAI)	NA	Acc 91.72%; Sens 98.45%; Spec 85.06%; F1 92.19%	Missing data handling not detailed; no real-world validation
Kang et al. [46]	7 University Hospitals, South Korea	34,387	Demographics, BMI, BP, labs, history	XGBoost (SHAP + Boruta)	70/30	AUC 0.804; AUPR 0.442	Retrospective bias; internal validation only
Zhang et al. [47]	Women's Hospital, Zhejiang Univ., China	3268	OGTT values	Logistic Regression	NA	aOR 2.66 (Macrosomia)	Not a prediction model; single-center retrospective
Zaky et al. [31]	HMC, Qatar	138	Clinical + biochemical biomarkers	Stacking (SHAP)	80/20	Acc 88.83%; F1 89.56%	Small, single-center; costly biomarkers
Cubillos et al. [48]	Hospital Parroquial, Chile	1611	Clinical, BMI, fasting glucose, parity	SVM / MLP (No XAI)	70/10/20	AUC 0.81–0.82	Single-center; no advanced biomarkers
Belsti et al. [44]	Monash Health, Australia	48,502	Demographic + obstetric history	CatBoost (FI)	80/20 + CV	AUC 0.93; Acc 85%; F1 81%	Single health network; no genetic biomarkers
Zhou et al. [30]	Two hospitals, Shenzhen, China	2309	Demographic, clinical, blood, social	XGBoost (SHAP)	80/20 + CV	AUC 0.913; Acc 85.0%	Regional restriction; biomarker inconsistency
Watanabe et al. [43]	JECS, Japan	82,698	Lifestyle, anthropometry, labs, SES	GBDT (SHAP)	80/20	AUC 0.74 (new), 0.67 (recurrent)	Low GDM incidence; single country
Hassan et al. [29]	Anbu Hospital & Nalam Clinic, India	3525	Clinical, OGTT, lifestyle, PCOS, family history	Fusion Ensemble (SHAP + LIME)	70/30 + CV	Acc 98.21%; F1 97.59%; AUC 99.91%	Single-region; no external clinical validation
Proposed	**Anbu Hospital & Nalam Clinic + 3 Kerala hospitals**	**3525 + 80**	**16 clinical features**	**TabNet (Attention + SHAP + LIME)**	**5-fold CV + external**	**Acc 97.13%±0.41; F1 96.22%±0.50; External F1 83.7–87.0%**	**Public training data; limited prospective external sample**

3.5. Explainability analysis

To ensure clinical transparency and trust, the proposed model integrates three complementary explainability techniques: built-in feature masks of TabNet, SHAP, and LIME. It is important to note that XAI-derived findings represent model-level feature attributions and interactions, and should not be interpreted as causal clinical relationships without further prospective clinical investigation.

3.5.1. Global feature importance

Global feature importance derived from TabNet masks is provided in Supplementary Figure S1. Prediabetes emerged as the top predictor, followed by PCOS, family history, and BMI—all consistent with established GDM risk factors reported in the clinical literature. OGTT, large child or birth defect, and unexplained prenatal loss also showed significant global weights, while diastolic BP, number of pregnancies, and haemoglobin were relatively less important.

The SHAP beeswarm plot (Fig. 4) confirms these model-level findings. High values of PCOS, prediabetes, and family history are associated with stronger model predictions toward the GDM class. BMI shows a clear dose-response pattern in model output, while Sys BP and OGTT have moderate model-level influence.

3.5.2. Feature interactions

The SHAP dependency plot for PCOS coloured by prediabetes status is provided in Supplementary Figure S2. When PCOS and prediabetes co-occur, the SHAP value rises strongly (up to 0.6+), indicating a strong model-identified interaction beyond either factor alone. This pattern is

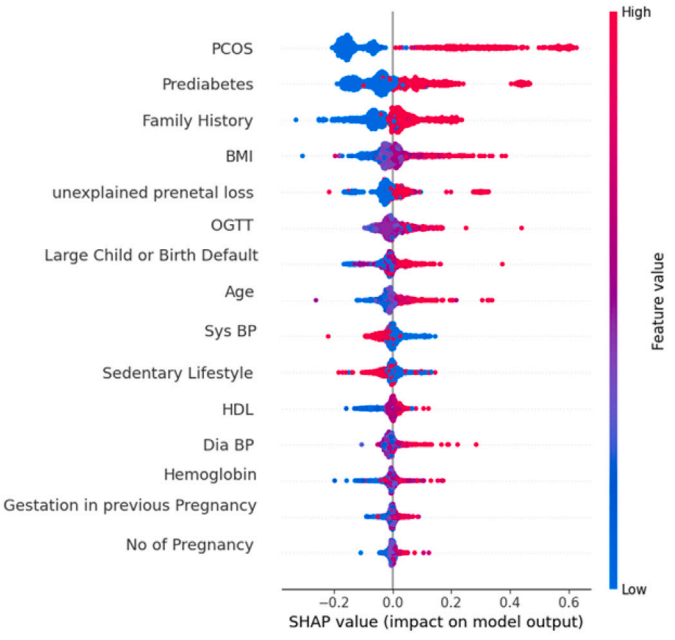


Fig. 4. SHAP beeswarm summary plot showing feature impact on model output across all samples. Colour indicates feature value; horizontal position indicates SHAP value. (For interpretation of the references to Colour in this figure legend, the reader is referred to the web version of this article.)

consistent with known clinical risk factor co-occurrence in GDM, though the interaction represents a model-level observation rather than a confirmed causal relationship. Supplementary Table S9 summarises this interaction. Similar amplification was observed for PCOS co-occurring with high BMI or positive family history.

3.5.3. Patient-level interpretability

TabNet local feature importance for two representative patients is provided in Supplementary Figure S3. For Instance 0, BMI, family history, and prediabetes dominated the model's decision; for Instance 1, age and OGTT were most influential—demonstrating instance-specific feature attribution.

The SHAP waterfall plot (Supplementary Figure S4) shows that prediabetes (+0.20), family history (+0.10), and gestation in previous pregnancy (+0.09) cumulatively shifted one patient's model output from baseline (0.49) toward a GDM prediction.

The LIME explanation (Supplementary Figure S5) assigns probability 1.00 to GDM driven by family history, prediabetes, and previous gestational issues—independently corroborating TabNet and SHAP model-level attributions.

Together, all three XAI methods consistently identify the same dominant model-level risk factors, suggesting robust and stable feature attribution across complementary explainability frameworks.

3.6. Physician-in-the-loop clinical validation

Four certified obstetricians (each >10 years of GDM experience) independently reviewed 30 stratified patient cases following established PITL protocols [49,50]. Each received all 16 features, the model's binary prediction, and top XAI explanations from TabNet, SHAP, and LIME. Reviewers remained blinded to true labels; majority voting determined consensus.

The model achieved 96.7% agreement with clinician consensus (29/30 cases). It should be noted that this PITL evaluation, while encouraging, is preliminary in nature due to the limited sample size (n = 30) and should not be interpreted as definitive evidence of clinical trustworthiness. Case 11 was the sole misclassification (model: Non-GDM; all four physicians: GDM), with no dominant SHAP feature, suggesting a complex multi-feature interaction not captured by the model. This false negative highlights the clinical importance of uncertainty-aware prediction and human override mechanisms in deployment settings. Case 20 showed one physician disagreement resolved by majority vote in line with the model. Full case-wise decisions are provided in Supplementary Table S10.

Cohen's kappa between the model and individual clinicians ranged from 0.856 to 0.927 (average 0.909), and Fleiss' kappa was 0.963 (Table 5), indicating strong physician–model and inter-rater agreement. These results are promising as a preliminary feasibility assessment, with larger-scale PITL evaluation recommended before drawing conclusions about clinical trustworthiness.

3.7. External multi-site clinical validation

To evaluate real-world generalizability, external validation was conducted across three independent hospitals in Kerala, India: Pariyaram Medical College (30 patients), SH Hospital (20 patients), and Aster Medcity (30 patients), comprising 80 prospectively collected patient records. Feature definitions were consistent with the training dataset. Supervising physicians confirmed the clinical diagnoses and reviewed model predictions for clinical relevance; feature importance outputs were examined to assess alignment with clinical expectations. Table 6 presents the comprehensive site-wise results. The TabNet classifier achieved an average F1 score of 85.69% across all sites, with performance ranging from 83.70% to 87.00%. The observed gap between internal validation performance (F1: 96.22%) and external validation performance (F1: 83.70–87.00%) is acknowledged as an expected outcome of applying a model trained on a single-source public dataset to prospectively collected data from independent clinical sites with inherent demographic and procedural variability. This gap underscores the need for multi-site training data in future model iterations, while confirming that the model maintains clinically reasonable predictive performance across diverse real-world settings.

Table 5
Agreement of TabNet model with Doctors and Inter-Doctor Reliability.

Doctor / Group	Disagreements	Cohen's κ
Doctor 1 vs Model	1 (Case 11)	0.927
Doctor 2 vs Model	1 (Case 11)	0.927
Doctor 3 vs Model	1 (Case 11)	0.927
Doctor 4 vs Model	2 (Cases 11 & 20)	0.856
Average Kappa	–	**0.909**
All 4 Doctors (Fleiss)	1 partially mixed	0.963

Table 6
Hospital-wise performance of the TabNet model.

Hospital	GDM	Non	Total	F1 (%)
Pariyaram Medical College	20	10	30	83.70
SH Hospital	9	11	20	86.37
Aster Medcity	14	16	30	87.00
Average			**80**	**85.69**

4. Discussion

The proposed framework integrates feature-tailored hybrid imputation, SMOTE-based class balancing, TabNet classification, and dual-stage clinical validation to address key challenges in GDM prediction. Feature-tailored imputation and SMOTE balancing improved data quality and class representation, contributing to stronger TabNet performance. TabNet achieved 97.13% accuracy and 98.91% recall, outperforming ten baseline classifiers under identical preprocessing conditions. All three XAI methods consistently identified prediabetes, PCOS, family history, and BMI as the dominant model-level risk factors, findings that are consistent with established clinical knowledge of GDM risk.

It is important to acknowledge, however, that the strong internal performance may partly reflect the homogeneity of the single-source training dataset and the complexity of the preprocessing pipeline. The use of SMOTE introduces synthetic samples that may not fully represent the true clinical distribution of GDM-positive cases, potentially contributing to optimistic internal performance estimates. The observed gap between internal validation (F1: 96.22%) and external validation (F1: 83.70–87.00%) is consistent with this interpretation and underscores the importance of treating internal metrics as upper-bound estimates rather than definitive measures of real-world performance. A key practical factor contributing to this performance reduction is the variability in feature availability and recording practices across the three independent Kerala hospitals. Different hospitals followed different clinical protocols, resulting in inconsistencies in how certain features were collected, measured, and recorded. This inter-site feature heterogeneity represents a significant real-world challenge for deploying standardised ML models across diverse clinical settings, and highlights the need for feature-harmonisation pipelines in future work. Against ten state-of-the-art studies, the proposed model matches or exceeds performance under a more rigorous 5-fold cross-validation strategy, while uniquely combining multi-site external validation with formal physician evaluation—a combination not previously reported for GDM prediction models.

SHAP-identified feature interactions. SHAP analysis identified a strong model-level interaction between PCOS and prediabetes (SHAP value > 0.6 at co-occurrence), independently corroborated by all four obstetricians. This pattern is consistent with known clinical co-occurrence of these risk factors in GDM. However, it is important to note that SHAP

quantifies feature contributions to model predictions and does not establish causal clinical relationships. These model-identified interactions should be interpreted as hypothesis-generating observations warranting further prospective clinical investigation rather than confirmed causal findings.

Clinical trustworthiness and PITL limitations. Physician-in-the-loop concordance of 96.7% ($\kappa_{\text{avg}} = 0.909$, $\kappa_F = 0.963$) is an encouraging preliminary result. However, clinical trustworthiness cannot be equated with agreement in a small PITL experiment involving 30 cases and four physicians. This evaluation should be regarded as a feasibility assessment demonstrating the plausibility of physician–model agreement, rather than definitive validation of clinical reliability. Larger-scale PITL studies involving diverse physician groups, varied patient populations, and prospective real-world deployment conditions are necessary before stronger claims of clinical trustworthiness can be made.

False negatives and clinical safety. The false negative in Case 11 (model prediction: Non-GDM; all four physicians: GDM) deserves particular attention, as false negatives carry serious clinical consequences in GDM. An undetected GDM case may result in unmanaged maternal hyperglycemia, increased risk of macrosomia, neonatal hypoglycemia, preeclampsia, and long-term metabolic complications for both mother and child. Case 11 showed no dominant SHAP feature, suggesting a complex multi-feature interaction that the model failed to capture. This highlights the critical need for uncertainty-aware prediction mechanisms, such as conformal prediction intervals, that can flag low-confidence cases for mandatory clinical review rather than allowing automated decisions. Any clinical deployment of this or similar models must incorporate explicit human override mechanisms, particularly for borderline or low-confidence predictions.

Real-world implementation challenges. Several practical challenges must be addressed before this framework can be integrated into clinical screening workflows. First, workflow integration requires that the model be embedded within existing electronic health record (EHR) systems in a manner that is minimally disruptive to clinical practice, with predictions presented alongside XAI explanations in a format interpretable by non-specialist clinicians. Second, decision threshold calibration is critical: the default 0.5 classification threshold may not be optimal for clinical use, where higher sensitivity (recall) is generally preferred to minimize false negatives, and site-specific threshold adjustment may be necessary. Third, ongoing model monitoring and periodic recalibration will be required to detect and correct for distributional shifts as patient populations, clinical protocols, and data recording practices evolve over time. Fourth, the current model was trained on a single-region dataset, and its performance in ethnically and geographically diverse populations remains to be established.

Limitations and future directions. Limitations of this work include: (1) single-region training data constraining ethnic and demographic diversity; (2) inter-site feature heterogeneity arising from differences in clinical protocols and feature recording practices across hospitals, which contributed to the observed performance reduction from internal validation (F1: 96.22%) to external validation (average F1: 85.69%); (3) a modest external cohort ($n = 80$) that precludes subgroup analysis by age, parity, or comorbidity; (4) the use of SMOTE-generated synthetic samples that may not fully reflect real clinical distributions; and (5) the geographically constrained external validation limited to Kerala, India. The false negative in Case 11 further exposes the need for uncertainty quantification mechanisms in clinical deployment. Future work will address these limitations through: expanded multi-ethnic and multi-site cohort collection with standardised feature recording protocols; development of feature-harmonisation pipelines to handle missing or inconsistently recorded variables across hospitals; conformal prediction for individualised uncertainty quantification; cost-sensitive learning as an alternative to SMOTE; calibration plot-based threshold optimization; and EHR integration toward MLTRL 6 deployment readiness [51].

5. Conclusion

This study presents an explainable TabNet framework with feature-tailored hybrid imputation and SMOTE-based class balancing for GDM risk assessment using routine first-antenatal-visit clinical variables. From a technical performance perspective, the model achieved 97.13% accuracy, 98.91% recall, and 96.22% F1-score (±0.50) via 5-fold cross-validation, outperforming ten baseline classifiers under identical preprocessing conditions.

From an interpretability perspective, TabNet masks, SHAP, and LIME consistently identified prediabetes, family history, PCOS, and BMI as the dominant model-level risk factors. SHAP dependency analysis revealed a strong model-identified interaction between PCOS and prediabetes, consistent with established clinical knowledge, though prospective clinical investigation is required before causal conclusions can be drawn.

From a clinical utility perspective, physician-in-the-loop validation yielded strong preliminary agreement ($\kappa_{\text{avg}} = 0.909$, $\kappa_F = 0.963$) across four experienced obstetricians, and multi-site external validation (F1: 83.70%–87.00%) demonstrated reasonable real-world performance across three independent hospitals. The observed performance gap between internal and external validation is partly attributed to inter-site feature heterogeneity arising from differences in clinical data recording protocols across hospitals—a key challenge for real-world deployment that future work will address through standardised feature harmonisation. These results are best interpreted as a clinically oriented proof-of-concept, demonstrating the feasibility of interpretable, physician-endorsed GDM screening using routine variables. Large-scale, multi-ethnic prospective validation with standardised data collection protocols remains a necessary prerequisite before clinical deployment can be considered.

CRediT authorship contribution statement

Anju Narayanan: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Praveen Sankaran:** Supervision, Project administration, Investigation, Funding acquisition, Formal analysis. **Uma V. Sankar:** Writing – review & editing, Resources, Funding acquisition. **Simi Kurian:** Writing – review & editing, Resources, Data curation. **Teslin John:** Writing – review & editing, Data curation.

Ethics approval

This study involved retrospective analysis of publicly available de-identified data and prospective collection of patient records under clinical supervision. All procedures involving human participants were performed in accordance with the Declaration of Helsinki. Written informed consent was obtained from all prospective participants prior to data collection. Ethical approval for patient recruitment at Aster Medcity, Kochi, was granted by the Institutional Ethics Committee (IEC), Aster Medcity (EC Registration No: ECR/737/ns/KL/2015/RR-21; Ref No: AM/EC/367-2023; approval dated 23 December 2023). As the study involved multiple centres, approvals from the respective governing bodies of the participating hospitals were obtained as required. All patient data were anonymised prior to computational analysis to ensure confidentiality and privacy.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Claude (Anthropic) for language refinement and formatting assistance, and Writefull (integrated with Overleaf) for grammar and language editing support. The authors reviewed and edited all content and take full responsibility for the published article.

Funding

This work is partially funded by the Indian Council of Medical Research (ICMR) under Project No. IIRPIG-2023-0001525, titled 'Effectiveness of a community-based peer-led P-MED intervention in Gestational Diabetes Mellitus (GDM) mothers of the southern state of India: A cluster randomized controlled pragmatic trial.'

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors sincerely thank the medical staff at all three participating hospitals for their valuable clinical support and for facilitating the prospective data collection during the external validation study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.ijmedinf.2026.106488.

Data availability

All data and material access requests can be forwarded to the corresponding author.

References

[1] L.A. Barbour, C.E. McCurdy, T.L. Hernandez, J.P. Kirwan, P.M. Catalano, J.E. Friedman, Cellular mechanisms for insulin resistance in normal pregnancy and gestational diabetes, Diabetes Care 30 (Supplement 2) (2007) S112–S119, https://doi.org/10.2337/dc07-s202

[2] H.D. McIntyre, P. Catalano, C. Zhang, M. Hod, J. Harreiter, M.A. Hirst, Gestational diabetes mellitus, Nat. Rev. Dis. Primers 5 (1) (2019) 47, https://doi.org/10.1038/s41572-019-0098-8

[3] L. Bellamy, J.P. Casas, A.D. Hingorani, D.A. Williams, Type 2 diabetes after gestational diabetes: a systematic review and meta-analysis, Lancet 373 (9677) (2009) 1773–1779, https://doi.org/10.1016/S0140-6736(09)60731-5

[4] B.E. Metzger, L.P. Lowe, A.R. Dyer, E.R. Trimble, U. Chaovarindr, D.R. Coustan, D.R. Hadden, D.R. McCance, M. Hod, H.D. McIntyre, J.J. Oats, B. Persson, M.S. Rogers, D.A. Sacks, HAPO Study Cooperative Research Group, Hyperglycemia and adverse pregnancy outcomes, New England Journal of Medicine 358 (19) (2008) 1991–2002, https://doi.org/10.1056/NEJMoa0707943

[5] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324

[6] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, https://doi.org/10.1007/BF00994018

[7] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539

[10] D. Farrar, L. Duley, T. Dowswell, D.A. Lawlor, Different strategies for diagnosing gestational diabetes to improve maternal and infant health, Cochrane Database Syst. Rev. 8 (2017) CD007122, https://doi.org/10.1002/14651858.CD007122.pub4

[11] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2 ed, John Wiley & Sons, 2002.

[12] A. Donders, G. van der Heijden, T. Stijnen, K. Moons, Review: a gentle introduction to imputation of missing values, J. Clin. Epidemiol. 59 (10) (2006) 1087–1091.

[13] G.E. Batista, M.C. Monard, An analysis of four missing data treatment methods for supervised learning, Appl. Artif. Intell. 17 (5–6) (2003) 519–533.

[14] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, 1987.

[15] D.J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (1) (2012) 112–118.

[16] P.C. Austin, I.R. White, D.S. Lee, S. van Buuren, Missing data in clinical research: a tutorial on multiple imputation, Can. J. Cardiol. 37 (9) (2021) 1322–1331, https://doi.org/10.1016/j.cjca.2020.11.010

[17] J. Yoon, J. Jordon, M. van der Schaar, GAIN: missing data imputation using generative adversarial nets, in: Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018, pp. 5675–5684.

[18] A. Kuang, Y. Yu, J. Siddique, D. Scholtens, Imputation of missing continuous glucose monitor data, J. Diabetes Sci. Technol. (2025), https://doi.org/10.1177/19322968241308217

[19] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953

[20] M. Hayaty, S. Muthmainah, S.M. Ghufran, Random and synthetic over-sampling approach to resolve data imbalance in classification, Int. J. Artif. Intell. Res. 4 (2) (2020) 86–94.

[21] I. Dey, V. Pratap, A comparative study of SMOTE, borderline-SMOTE, and ADASYN oversampling techniques using different classifiers, in: 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), IEEE, 2023, pp. 294–302.

[22] S. Akbar, A. Raza, W. Alghamdi, H. Ali, Q. Zou, X. Luo, Identifying protein succinylation sites using generative transformer and a two-dimensional representation with a deep capsule network, iScience 28 (12) (2025).

[23] S. Akbar, A. Raza, W. Alghamdi, A. Saeed, H. Ali, Q. Zou, DeepAIPs-SFLA: deep convolutional model for prediction of anti-inflammatory peptides using binary pattern decomposition of novel multiview descriptors with an SFLA approach, ACS Omega 10 (32) (2025) 35747–35762, https://doi.org/10.1021/acsomega.5c02422

[24] S. Akbar, A. Raza, H.H. Awan, Q. Zou, W. Alghamdi, A. Saeed, pNPs-CapsNet: predicting neuropeptides using protein language models and FastText encoding-based weighted multi-view feature integration with deep capsule neural network, ACS Omega 10 (12) (2025) 12403–12416.

[25] S.Ö. Arik, T. Pfister, TabNet: attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 6679–6687.

[26] M.N.H. Chowdhury, M.B.I. Reaz, S.H.M. Ali, et al., Deep learning for early detection of chronic kidney disease stages in diabetes patients: a TabNet approach, Artif. Intell. Med. 166 (2025) 103153.

[27] Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using Shapley additive explanation and application for real data in hospital, Comput. Methods Programs Biomed. 214 (2022) 106584.

[28] V. Vishwarupe, P.M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, V. Pawar, Explainable AI and interpretable machine learning: a case study in perspective, Procedia Comput. Sci. 204 (2022) 869–876.

[29] A. Hassan, S.G. Ahmad, T. Iqbal, E.U. Munir, K. Ayyub, N. Ramzan, Enhanced model for gestational diabetes mellitus prediction using a fusion technique of multiple algorithms with explainability, Int. J. Comput. Intell. Syst. 18 (1) (2025) 47, https://doi.org/10.1007/s44196-025-00760-4

[30] F. Zhou, X. Ran, F. Song, Q. Wu, Y. Jia, Y. Liang, S. Chen, G. Zhang, J. Dong, Y. Wang, A stepwise prediction and interpretation of gestational diabetes mellitus: foster the practical application of machine learning in clinical decision, Heliyon 10 (12) (2024) e32709, https://doi.org/10.1016/j.heliyon.2024.e32709

[31] H. Zaky, E. Fthenou, L. Srour, T. Farrell, M. Bashir, N. El Hajj, T. Alam, Machine learning based model for the early detection of gestational diabetes mellitus, BMC Med. Inform. Decis. Mak. 25 (1) (2025) 130, https://doi.org/10.1186/s12911-025-02576-2

[32] A. Sumathi, S. Meganathan, B.V. Ravisankar, An intelligent gestational diabetes diagnosis model using deep stacked autoencoder, Computers, Materials & Continua 69 (3) (2021) 3109–3126.

[33] L.O. Joel, W. Doorsamy, B.S. Paul, On the performance of imputation techniques for missing values on healthcare datasets, arXiv preprint, 2024 arXiv:2403.14687). https://arxiv.org/abs/2403.14687.

[34] C.M. Musil, C.B. Warner, P.K. Yobas, S.L. Jones, A comparison of imputation techniques for handling missing data, West. J. Nurs. Res. 24 (7) (2002) 815–829, https://doi.org/10.1177/019394502762477004

[35] R.S. Selina, M. Rahardi, A. Aminuddin, F.F. Abdulloh, H. Badi, B.P. Asaddulloh, Optimizing diabetes diagnosis using machine learning with SMOTE and feature selection, in: 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI), 2025, pp. 647–652, https://doi.org/10.1109/ICoCSETI63724.2025.11020043

[36] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks, 2008, pp. 1322–1328, https://doi.org/10.1109/IJCNN.2008.4633969

[37] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259, https://doi.org/10.1016/j.neunet.2018.07.011

[38] G. Naidu, T. Zuva, E.M. Sibanda, A review of evaluation metrics in machine learning algorithms, in: Computer Science on-Line Conference, Springer, 2023, pp. 15–25.

[39] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.

[40] M. Li, T. Yu, Methodological issues on evaluating agreement between two detection methods by cohen's kappa analysis, Parasites & Vectors 15 (1) (2022) 270.

[41] G. Rücker, T. Schimek-Jasch, U. Nestle, Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using fleiss' kappa, Methods Inf. Med. 51 (6) (2012) 489–494.

[42] Q. Ji, L. Gao, H. Liu, X. Chen, B. Fu, Y. Lin, F. Wang, Early prediction of gestational diabetes mellitus using machine learning-integrated metabolomic and clinical features, Front. Endocrinol. 16 (2025) 1687146, https://doi.org/10.3389/fendo.2025.1687146

[43] M. Watanabe, A. Eguchi, K. Sakurai, M. Yamamoto, C. Mori, The Japan Environment Children's Study (JECS) Group, Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan environment and children's study, Sci. Rep. 13 (1) (2023) 17419, https://doi.org/10.1038/s41598-023-44313-1

[44] Y. Belsti, L. Moran, L. Du, A. Mousa, K. De Silva, J. Enticott, H. Teede, Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the monash GDM machine learning model, Int. J. Med. Inform. 179 (2023) 105228, https://doi.org/10.1016/j.ijmedinf.2023.105228

[45] O.O. Abe, O.O. Obe, O.K. Boyinbode, N.B. Olagbuji, Early gestational diabetes mellitus diagnosis using classification algorithms: an ensemble approach, in: 2023 IEEE Africon, 2023, pp. 1–6, https://doi.org/10.1109/AFRICON55910.2023.10293366

[46] B.S. Kang, S.U. Lee, S. Hong, S.K. Choi, J.E. Shin, J.H. Wie, Y.S. Jo, Y.H. Kim, K. Kil, Y.H. Chung, K. Jung, H. Hong, I.Y. Park, H.S. Ko, Prediction of gestational diabetes mellitus in asian women using machine learning algorithms, Sci. Rep. 13 (1) (2023) 13356, https://doi.org/10.1038/s41598-023-39680-8

[47] Y. Zhang, L. Chen, Y. Ouyang, X. Wang, T. Fu, G. Yan, Z. Liang, D. Chen, A new classification method for gestational diabetes mellitus: a study on the relationship between abnormal blood glucose values at different time points in oral glucose tolerance test and adverse maternal and neonatal outcomes, AJOG Global Reports 4 (4) (2024) 100390, https://doi.org/10.1016/j.xagr.2024.100390

[48] G. Cubillos, M. Monckeberg, A. Plaza, M. Morgan, P.A. Estevez, M. Choolani, M.W. Kemp, S.E. Illanes, C.A. Perez, Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy, BMC Pregnancy Childbirth 23 (1) (2023) 469, https://doi.org/10.1186/s12884-023-05766-4

[49] L. Bigorra, I. Larriba, R. Gutiérrez-Gallego, A physician-in-the-loop approach by means of machine learning for the diagnosis of lymphocytosis in the clinical laboratory, Arch. Pathol. Lab. Med. 146 (8) (2022) 1024–1031, https://doi.org/10.5858/arpa.2021-0044-OA

[50] N.J. Bull, B. Honan, N.J. Spratt, S. Quilty, A method for rapid machine learning development for data mining with doctor-in-the-loop, PLOS ONE 18 (5) (2023) e0284965, https://doi.org/10.1371/journal.pone.0284965

[51] A. Lavin, C.M. Gilligan-Lee, A. Visnjic, S. Ganju, D. Newman, A.G. Baydin, et al., Technology readiness levels for machine learning systems, arXiv preprint, 2021 arXiv:2101.03989.