EYE-GDM: Clinically Validated, Explainable Ensemble Learning for Gestational Diabetes

Shatha Alghamdi[1], Rashid Mehmood[2], Fahad Alqurashi[3], Turki Alghamdi[4], Sarah Ghazali[5], Asmaa AlAhmadi[6]
Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[1,3]
Computer Science & Artificial Intelligence Department, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia[1]
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia[2,4]
Department of obstetrics and gynecology, College of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia[5]
Department of Women and Child Health (Obstetrics and Gynecology), College of Medicine, Taibah University, Madinah, Saudi Arabia[6]

***Abstract*—As artificial intelligence (AI) advances in healthcare, its use in maternal health shows promise but faces challenges of trust due to the black-box nature of many models. Gestational diabetes mellitus (GDM), a transient yet high-risk condition, demands accurate and interpretable prediction tools. However, existing GDM prediction studies often rely on opaque models or post-hoc explanation techniques applied after training, which limits transparency and reduces their clinical applicability. This highlights an urgent need for models that unify high predictive performance with interpretability by design. This study introduces EYE-GDM, a case-specific application of our Enhanced Interpretability Ensemble (EYE) framework, designed to predict GDM risk with clinically meaningful explanations. The pipeline evaluates multiple algorithms and selects Decision Tree (DT), k-Nearest Neighbors (k-NN), and Gradient Boosting (GB) as the best-performing base learners. These are integrated with SHAP and a logistic regression (LR) meta-model to construct EYE-GDM, embedding interpretability by weighting learner outputs with LR coefficients. This yields global (population-level) and local (patient-level) explanations consistent with medical knowledge. Tested on a dataset of 3,525 pregnancies, EYE-GDM achieved strong performance (accuracy = 0.9789, AUC-ROC = 0.9981) and provided insights into risk patterns, thresholds, and feature interactions relevant to GDM. By embedding explainability within the ensemble construction, EYE-GDM achieves transparent and clinically aligned reasoning without compromising predictive performance. Thus, EYE-GDM demonstrates how explainable AI (XAI) can translate from technical innovation to practical value in maternal care, supporting earlier risk identification and more informed clinical decisions.**

***Keywords*—Explainable Artificial Intelligence (XAI); interpretable machine learning (IML); Gestational diabetes mellitus (GDM); maternal health; healthcare AI; GDM risk prediction; transparency; trust**

I. Introduction

Artificial intelligence (AI) is increasingly being used in healthcare to support early risk detection and guide clinical decision-making [1]. Yet, despite its growing presence, many AI models, especially high-performing ones such as deep learning and ensemble methods, suffer from the black-box problem and remain difficult for clinicians to trust and use [2]. In areas such as maternal health, where model predictions can directly affect outcomes for both mother and child, transparency is not optional. Clinicians need more than just a prediction; they need a clear rationale they can understand and explain. In this context, integrating explainable AI (XAI) into decision support systems is essential, as it can significantly affect clinicians' trust and the extent to which they follow AI-driven recommendations [3].

Gestational diabetes mellitus (GDM) affects a substantial number of pregnancies worldwide and is linked to serious complications such as preeclampsia, macrosomia, neonatal hypoglycaemia, and future type 2 diabetes. Its transient nature and rapid physiological onset during pregnancy create a narrow window for timely risk identification and intervention [4]. In addition, GDM has been associated with adverse pregnancy outcomes including preterm birth, hypertensive disorders, shoulder dystocia, hyperbilirubinemia, stillbirth, and caesarean delivery [5], [6], [7]. Therefore, there is an urgent need for clinically interpretable machine learning methods tailored to GDM patients, so that predictions can be understood and applied meaningfully in practice [8].

Although machine learning (ML) has been widely applied to GDM prediction, many models rely on post-hoc explainability methods such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations). These approaches provide some insight but do so after the model has been trained, treating explanation as a separate process. This separation often limits clinical usefulness, as the reasoning behind predictions may not fully reflect the model's internal decision logic [9], [10]; see Section II for related work and research gap.

To address the dual challenge of predictive performance and interpretability, we previously introduced the Enhanced Interpretability Ensemble (EYE) framework [11]. EYE is a structured methodology that embeds interpretability throughout the model development pipeline, from defining the clinical application and preparing data to selecting base learners, integrating explanation methods, and constructing ensembles that remain transparent by design. Unlike post hoc approaches, EYE aligns interpretability with the model's internal reasoning.

We first implemented this framework in EYE-WD [11], focused on diabetes risk prediction in women, where it achieved competitive performance across multiple datasets while uncovering clinically consistent explanations. This work highlighted the value of building AI systems that capture health patterns shaped by gender-specific factors, ensuring that predictive tools serve populations often overlooked in women's and maternal health research.

Building on this foundation, the present work applies the EYE framework to gestational diabetes mellitus (GDM), a transient but high-risk condition unique to pregnancy, resulting in EYE-GDM. The ensemble integrates Decision Tree (DT), k-Nearest Neighbors (k-NN), and Gradient Boosting (GB) as base learners, combined through a logistic regression (LR) meta-model with SHAP explanations embedded in training. Applied to a dataset of 3,525 pregnancies, EYE-GDM achieved an accuracy of 97.89%, F1-score of 97.90%, precision of 97.47%, recall of 98.33%, and an AUC-ROC of 0.9981. Beyond these results, interpretability analyses produced clinically meaningful insights. Global SHAP patterns revealed thresholds such as elevated risk at BMI values above 25 and OGTT values exceeding 160 mg/dL, while HDL levels above 35 mg/dL were protective. Interactions between features, for example, high BMI with low HDL or multiple pregnancies with elevated blood pressure, revealed clusters of maternal risk factors. Local explanations further traced patient-specific outcomes, distinguishing protective from adverse profiles in a transparent way.

In doing so, this study advances GDM risk assessment and supports the broader goal of developing XAI that can be integrated into real-world maternal-care workflows [12]. By presenting EYE-GDM alongside the earlier EYE-WD implementation, we demonstrate both the flexibility and clinical relevance of the EYE framework and its potential to address critical challenges in women's and maternal health, providing contributions that are meaningful to technical experts and healthcare professionals alike.

The remainder of this study is organized as follows: Section II reviews the related work. Section III provides the methodology. Section IV reports the performance results of EYE-GDM, while Section V provides clinically oriented interpretations, including dataset-specific risk patterns, feature interactions, data-driven thresholds, and patient-level interpretations. Section VI provides a discussion, and Section VII concludes the study with directions for future work.

II. Related Work

GDM has received growing attention from both clinical and computational research communities. In particular, a substantial body of work has applied machine learning techniques to develop predictive models for GDM, aiming to support early diagnosis and timely intervention that can improve maternal and neonatal outcomes. Section II A reviews studies that develop machine learning and deep learning models to identify women at risk of GDM. Section II.B examines a growing body of work on making these predictions understandable to clinicians.

A. Machine Learning and Deep Learning for GDM

Sumathi et al. [13], [14] constructed a GDM dataset, which includes 3,525 pregnant women's data. Using deep learning, they introduced a deep stacked autoencoder based on outlier detection process (OD-DSAE), achieving 96.18% accuracy [13], and later proposed an ensemble voting strategy combining k-nearest neighbours (k-NN), Random Forest (RF), and LR models, reaching 94.24% accuracy [14]. Jader et al. [15] employed an ensemble approach using lab records from Iraq-Kurdistan, achieving 92% accuracy through majority voting with Decision Tree (DT), RF, Support Vector Machine (SVM), k-NN, LR, and Naïve Bayes (NB). Gallardo-Rincón et al. [16] developed and evaluated an Artificial Neural Network (ANN) model, achieving 70.3% accuracy in recognizing women at high risk of developing GDM. This model was based on data from 1,709 pregnant Mexican women who participated in the 'Cuido mi embarazo' study. Zheng et al. [17] collected data on 4,771 Chinese pregnant women from Xinhua Hospital. They applied Multivariate LR using Bayesian inference and achieved an accuracy of 64% and an AUC of 0.766. Shen et al. [18] evaluated several machine learning algorithms, including LR, SVM, RF, AdaBoost, DT, NB, k-NN, XGBoost, and Gradient Boosting Decision Tree (GBDT), to predict GDM in areas with limited resources. The algorithms were trained on 12,304 cases from the First Affiliated Hospital of Jinan University and were validated on 1,655 pregnant patients at the Prince of Wales Hospital. At 0.78 AUC, SVM showed the best performance.

Wu et al. [19] et al. utilized LR, k-NN, SVM, and deep neural network (DNN) for early GDM prediction. They evaluated these machine learning models using data from the International Peace Maternal and Child Health Hospital at Shanghai Jiao Tong University School of Medicine, which included 32,190 pregnant patients. DNN obtained the best AUC of 0.8, followed by LR with an AUC of 0.77. For predicting GDM, Hu et al. [20] utilized and compared between traditional LR method and the XGBoost machine learning model on a total of 925 pregnant women. They demonstrated that XGBoost outperformed LR with an AUC of 0.946 and an accuracy of 87.5%. Through the Japan Environment and Children's Study (JECS), Watanabe et al. [21] gathered GDM data from 82,698 expectant women. They employed LR, SVM, GBDT, and RF for gestational diabetes detection. They demonstrated that GBDT produced the best AUC of 0.74.

Table I summarizes machine learning and deep learning research used to predict women at risk of GDM. The first two columns provide the reference number and a summary of each work. The third column presents the reported performance, while the fourth column records limitations and key observations.

TABLE I. SUMMARY OF WORKS ON GDM PREDICTION

Ref	Work Summary	Performance	Limitations / Observations
[13]	Developed a deep learning model using a stacked autoencoder with outlier detection (OD-DSAE).	Accuracy: 96.18%	• Lack interpretability
[14]	Constructed an ensemble voting model using k-NN, RF, and LR.	Accuracy: 94.24%	
[15]	Trained and evaluated an ensemble of six classifiers (DT, RF, SVM, k-NN, LR, NB) on data from Iraq-Kurdistan using majority voting.	Accuracy: 92%	
[16]	Evaluated the ANN model on data from 1,709 pregnant Mexican women.	Accuracy: 70.3%	• Low performance • Lack interpretability
[17]	Applied Multivariate LR using Bayesian inference on 4,771 Chinese pregnant women from Xinhua Hospital.	Accuracy: 64%, AUC: 0.766	
[18]	Compared nine ML models (LR, SVM, RF, AdaBoost, DT, NB, k-NN, XGBoost, and GBDT) trained on 12,304 cases and were validated on 1,655 pregnant patients.	AUC: 0.78 (SVM best)	
[19]	Evaluated LR, k-NN, SVM, and DNN on the Shanghai dataset (32,190 pregnant patients).	AUC: 0.8 (DNN best)	• Moderate performance • Lack interpretability
[20]	Compared LR and XGBoost on 925 pregnant women.	Accuracy: 87.5% (XGB best)	
[21]	Assessed a large-scale Japanese cohort (82,698 pregnant women) with four models LR, SVM, GBDT, and RF.	AUC: 0.74 (GBDT best)	• Low performance • Lack interpretability

B. Explainability and Interpretability in GDM Prediction

Despite these advances in predictive performance, most previous works did not incorporate interpretability into their predictive models. To the best of our knowledge, only a limited number of studies have explicitly integrated explainability. For instance, Du et al. [22] applied SHAP with multiple machine learning models (AdaBoost, LR, SVM, RF, XGBoost) and used Synthetic Minority Oversampling (SMOTE) to balance the data. They reported an accuracy of 76.1%, with SVM performing the best. Khanna et al. [23] evaluated several machine learning pipelines and data balancing strategies for GDM prediction, where a stacking ensemble trained on SMOTE-ENN (Synthetic Minority Oversampling using Edited Nearest Neighbour) data achieved the best performance with an accuracy of 96% and an AUC of 0.96. They applied five post-hoc methods, including SHAP, LIME, ELI5, Qlattice, and Anchor, which confirmed the links between visceral fat levels, the child's birth weight, and GDM.

Zaky et al. [24] explored a range of algorithms, from traditional models such as LR and DT to advanced ensembles such as RF, Gradient Boosting (GB), CatBoost, XGBoost, and LightGBM, and then combined them in a stacking framework with LR as the meta-classifier. The stacked model trained on 26 selected features gave the best results with an accuracy of 88.8%, outperforming all individual models. SHAP was applied as a post-hoc tool to interpret predictions and highlight key biomarkers. While these contributions have advanced GDM prediction, they largely treat explainability as an afterthought rather than a property embedded within the model itself.

Table II summarizes prior works on GDM prediction that incorporate explainability or interpretability. Column 1 lists the reference; Column 2 summarizes the work; Column 3 reports performance metrics; Column 4 notes limitations and observations.

C. Research Gap

Nevertheless, the studies that do incorporate explainability rely mainly on post-hoc methods such as SHAP and LIME, applied only after the model has been trained. While these approaches offer valuable insights, they do not influence the model's internal structure or learning process, so interpretability remains an add-on rather than a built-in property.

Our overarching aim is to deliver accurate and transparent decision support in high-stakes domains by embedding interpretability across the full modelling workflow, aligning design and evaluation with stakeholder requirements to strengthen accountability, and enabling real-world use without sacrificing predictive performance. We previously introduced the FIXAIH framework [25], which articulates six well-defined design priorities to guide the development of trustworthy, interpretable AI for healthcare. Within this vision, we also proposed the EYE framework [11], which operationalizes selected FIXAIH priorities into a concrete architecture, providing interpretability by design, clinical validation, and transparent model development. We initially implemented EYE for diabetes risk in women (EYE-WD).

In this work, we extend EYE to GDM and refer to this implementation as EYE-GDM. Our approach integrates SHAP within the training pipeline, combining base-model outputs via meta-model weights to produce both local and global explanations that reflect the model's internal structure and how predictions are formed, while maintaining strong predictive performance by employing top-performing base learners during model construction. This work improves GDM risk assessment while contributing to the wider aim of building XAI systems suitable for integration into routine maternal-care workflows.

TABLE II. SUMMARY OF RELATED WORKS ON GDM PREDICTION INCORPORATING EXPLAINABILITY

Ref	Work Summary	Performance	Limitations / Observations
[22]	Applied SHAP with multiple ML models: AdaBoost, LR, SVM, RF, and XGBoost with SVM performing the best.	Accuracy: 76.1%	Post-hoc explainability only; not embedded
[23]	Evaluated multiple ML pipelines and data balancing strategies. A stacking ensemble trained on SMOTE-ENN gave the best results. SHAP, LIME, ELI5, Qlattice, and Anchor were applied to interpret results.	Accuracy: 96%, AUC: 0.96	
[24]	Explored a wide range of ML models (LR, DT, NB, RF, GB, CatBoost, XGBoost, LIGHTGBM), combined in a stacking framework with LR as meta-classifier. SHAP was used to explain predictions and rank biomarkers.	Accuracy: 88.8%	
EYE-GDM	Constructed an interpretable ensemble using EYE framework, integrating top-performing base learners, SHAP tool, and LR as meta-model. SHAP is applied internally during training, combining base learner outputs via meta-model weights to generate explanations.	Accuracy: 97.89%, AUC: 0.99	• Implementation of EYE framework • Embedded interpretability • Competitive performance

III. METHODOLOGY AND DESIGN

In this section, we present the methodology and design of the EYE-GDM system. Owing to space limitations, only a summary is provided here. A detailed description of the methodology, the EYE framework, and its initial implementation in EYE-WD can be found in [11]. EYE-GDM builds upon this foundation by extending the EYE framework into a domain-specific configuration tailored for gestational diabetes. The methodological refinements include optimizing the selection of ensemble base learners for pregnancy-related data characteristics, applying explanation-weighted meta-learning to integrate model reasoning with prediction, and aligning interpretability evaluation with pregnancy-specific clinical variables and thresholds. These adaptations expand the framework's methodological scope while preserving its interpretability-by-design foundation. Fig. 1 shows the architecture of the EYE-GDM system, which consists of eight components.

The Application Selection (APS) targets GDM risk prediction. In Data Preparation and Management (DPM), we use a publicly available dataset of 3,525 patients described by 15 clinical and demographic features, labelled into two classes: 2,153 non-GDM and 1,372 GDM cases. The dataset was collected by researchers at SASTRA Deemed to be University, Kumbakonam, Tamil Nadu, India, in collaboration with a consultant obstetrician and gynaecologist from Anbu Hospital and Nalam Clinic. It is publicly available on Kaggle under the CC BY-NC-SA 4.0 license [26] and was previously used in [13] and [14] to develop and evaluate GDM prediction models. The features are Age, Body Mass Index (BMI), Hemoglobin, Systolic Blood Pressure (Sys BP), Diastolic Blood Pressure (Dia BP), Oral Glucose Tolerance Test (OGTT), Prediabetes, Polycystic Ovary Syndrome (PCOS), Family History, Sedentary Lifestyle, Number of Pregnancies, Large Child or Birth Defect, Gestation in Previous Pregnancy, Unexplained Prenatal Loss, and High-Density Lipoprotein (HDL) (see Table III). All features are continuous or binary (0/1); binaries use 0 for absence and 1 for presence. We apply class-wise imputation for missing values, treat outliers using interquartile range (IQR) rules, scale features, and create a stratified train-test split to preserve class distribution.

Base-Model Evaluation and Selection (BMES) evaluates LR, Ridge Classifier (RC), DT, k-NN, NB, RF, AdaBoost, GB, Histogram Gradient Boost (HGB), and XGBoost on the training split using stratified k-fold cross-validation with hyperparameter tuning. We record AUC-ROC, F1-score, precision, recall, and accuracy, then select the top-performing models as the base learners for EYE-GDM.

Interpretable Meta-Model Selection (IM2S) uses LR as the stack combiner because it provides readable coefficients, well-calibrated probability scores, and efficient training and inference [27]. Explainability Tool Selection (XTS) adopts SHAP since it is model-agnostic, grounded in game-theoretic principles, supports dataset-level and patient-level views, and explains how features move a prediction from a baseline [28], [29].

TABLE III. DESCRIPTION OF FEATURES IN GDM DATASET

Feature	Description
Age	Maternal age at time of pregnancy
BMI	Body Mass Index
Hemoglobin	Hemoglobin level
Sys BP	Systolic Blood Pressure
Dia BP	Diastolic Blood Pressure
OGTT	Oral Glucose Tolerance Test result
Prediabetes	History of prediabetes
PCOS	Polycystic Ovary Syndrome diagnosis
Family History	Family history of diabetes
Sedentary Lifestyle	Lack of regular physical activity
No of Pregnancy	Number of prior pregnancies
Large Child or Birth Default	Previous large baby or complicated delivery
Gestation in Previous Pregnancy	Prior gestational diabetes or complications
Unexplained prenatal loss	History of miscarriage or fetal loss without known cause
HDL	High-Density Lipoprotein cholesterol

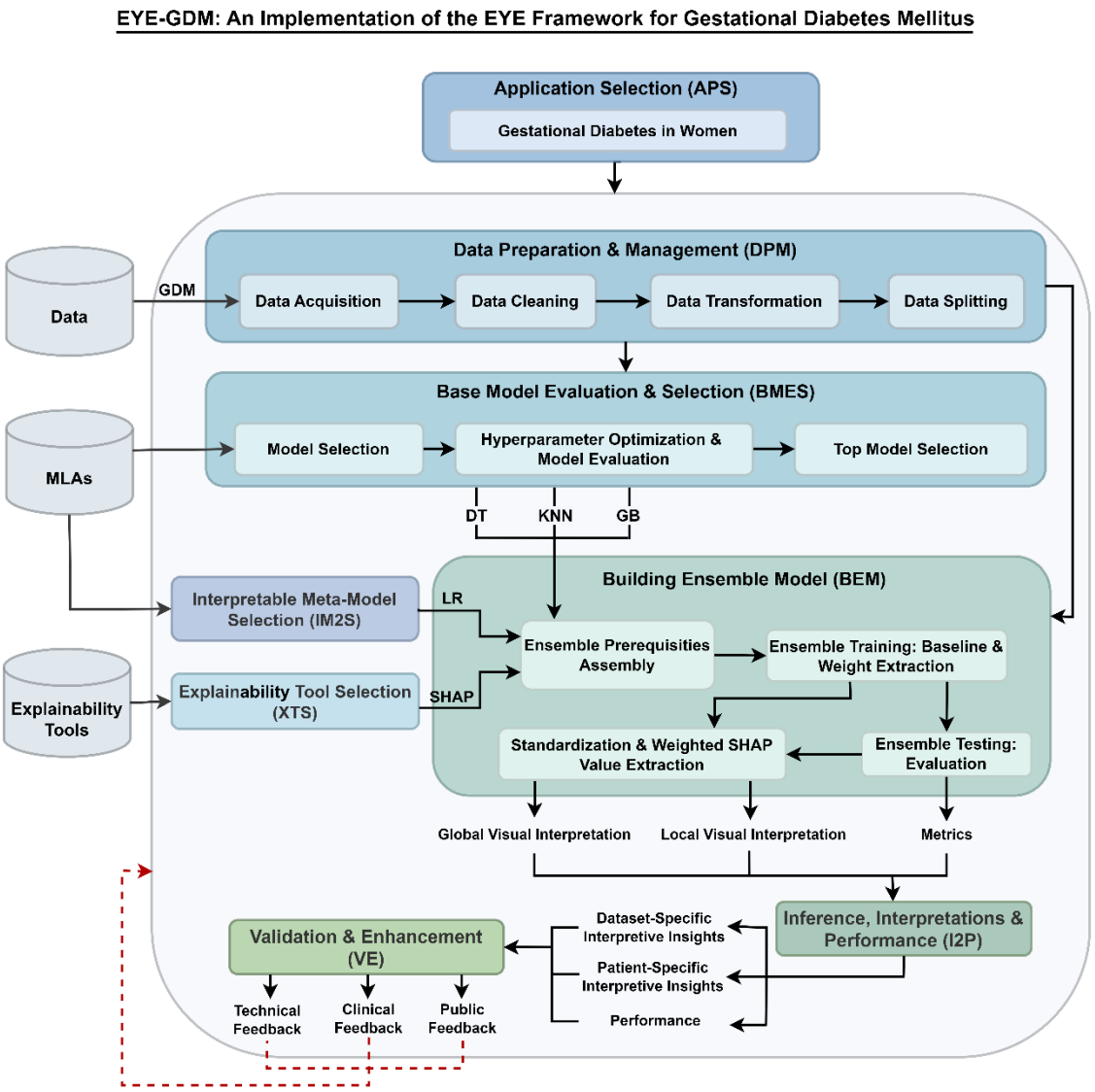


Fig. 1. The EYE-GDM System: An implementation of the EYE framework for gestational diabetes.

Building the Ensemble (BEM) constructs the interpretable ensemble. During training, each base learner yields SHAP values and expected values. We standardize these values to make magnitudes comparable, then form a single explanation by weighing each model's SHAP vector with the learned LR coefficients. This produces consolidated local and global explanations aligned with the ensemble's decision rule. Within BEM, these explanations are rendered as SHAP summary plots (global interpretations), dependence plots (pairwise effects and thresholds), and waterfall plots (patient-level breakdown).

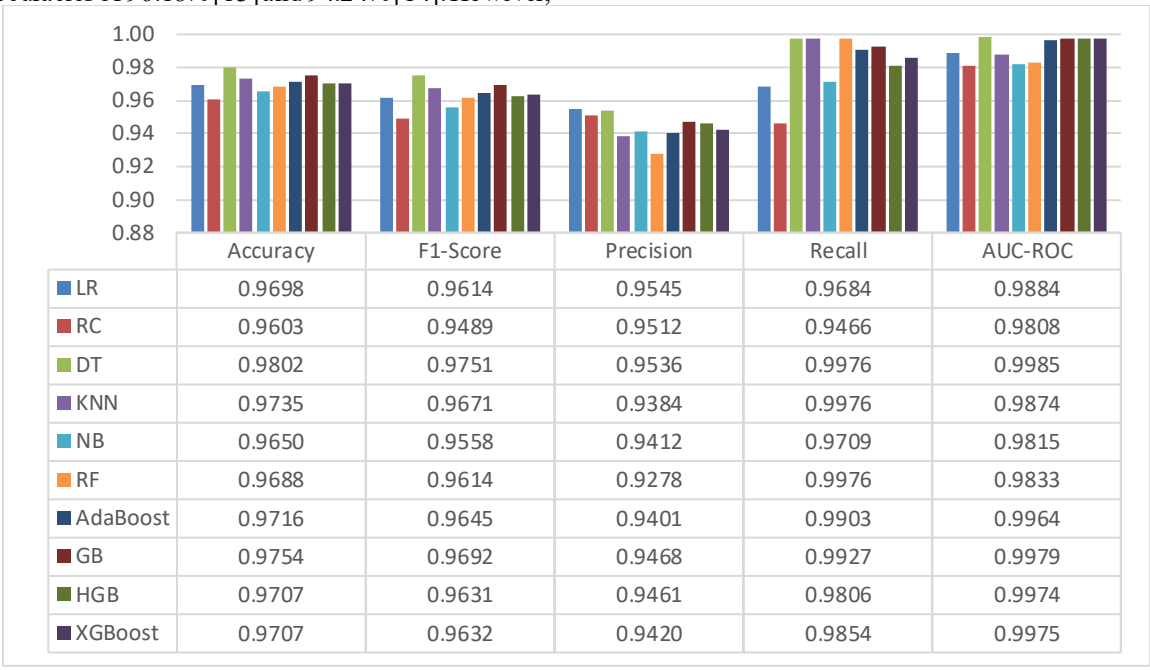
Inference, Interpretation, and Performance (I2P) uses these plots and prediction scores to deliver clinically relevant, fused explanations at both patient and dataset levels, clarifying which factors raise or lower risk, identifying data-driven thresholds, and characterizing feature interactions. Validation and Enhancement (VE) validates both technical and clinical outcomes: technical validation benchmarks performance against prior studies and includes review by technical experts, while clinical validation is conducted by clinical experts to assess alignment with medical standards. Their evaluation focused on adherence to medical knowledge, consistency with clinical practice, validity of diagnostic reasoning, clarity of explanations, and recommendations for improvement. Accepted adjustments are then applied and re-tested on the held-out split.

IV. EYE-GDM: Results

In this section, we present results from the BMES component, which evaluates multiple ML algorithms to identify the top-performing base learners for constructing EYE-GDM. Fig. 2 shows performance across five metrics (accuracy, F1, precision, recall, and AUC-ROC). DT, k-NN, and GB achieved the highest accuracies, with AUC-ROC values around 0.98, indicating clear separation between diabetic and non-diabetic cases. These three models were therefore selected as the ensemble's base learners. During the training phase, the LR meta-model determined contribution weights for each base

learner according to its influence on the overall ensemble output. The resulting coefficients were 6.96 for DT, 1.86 for k-NN, and 0.77 for GB, reflecting their relative influence in the final prediction. These weights were then applied to combine the base learners' explanation outputs, ensuring that the ensemble's interpretability aligns with its internal decision logic.

We evaluated EYE-GDM on the GDM dataset and compared its predictive performance with prior works that used the same data. As shown in Table IV, the reference models achieved accuracies of 96.18% [13] and 94.24% [14]. However, those studies reported only performance and did not provide interpretability or clinical explanations. In contrast, EYE-GDM achieved an accuracy of 97.89%, an F1-score of 97.90%, a precision of 97.47%, a recall of 98.33%, and an AUC-ROC of 0.9981. These results highlight that the framework shows competitive performance compared to previously reported baselines in predictive reliability. More importantly, it does so while also offering transparent and clinically meaningful explanations of its predictions, which prior models lacked. This dual capability of competitive performance and interpretability forms the basis for the subsequent analysis of GDM risk factors.



	Accuracy	F1-Score	Precision	Recall	AUC-ROC
LR	0.9698	0.9614	0.9545	0.9684	0.9884
RC	0.9603	0.9489	0.9512	0.9466	0.9808
DT	0.9802	0.9751	0.9536	0.9976	0.9985
KNN	0.9735	0.9671	0.9384	0.9976	0.9874
NB	0.9650	0.9558	0.9412	0.9709	0.9815
RF	0.9688	0.9614	0.9278	0.9976	0.9833
AdaBoost	0.9716	0.9645	0.9401	0.9903	0.9964
GB	0.9754	0.9692	0.9468	0.9927	0.9979
HGB	0.9707	0.9631	0.9461	0.9806	0.9974
XGBoost	0.9707	0.9632	0.9420	0.9854	0.9975

Fig. 2. Performance comparison across candidate base learners.

TABLE IV. PERFORMANCE COMPARISON OF EYE-GDM AGAINST PRIOR STUDIES ON THE SAME DATASET

Work	Accuracy	F1-score	Precision	Recall	AUC-ROC
[13]	0.9618	0.9741	0.9617	0.9869	-
[14]	0.9424	0.9400	0.9400	0.9400	-
EYE-GDM	0.9789	0.9790	0.9747	0.9833	0.9981

V. EYE-GDM: CLINICAL INTERPRETATIONS

In this section, we provide interpretations at both the dataset-specific and the individual patient-specific, alongside an analysis of feature interactions and dataset-specific thresholds. Different datasets, representing distinct populations, may yield unique thresholds influenced by lifestyle and other population-specific factors [30]. Therefore, the thresholds identified here should be regarded as dataset-specific, and validation across different populations is necessary. In Section V A, we use a SHAP summary plot to interpret global feature contributions to GDM risk prediction, thereby identifying dataset-specific interpretations. SHAP dependence plots are then applied in Section V B to capture feature interactions and define dataset-specific thresholds. Finally, in Section V C, we illustrate patient-level interpretation by selecting a representative patient case and applying a SHAP waterfall plot. In these SHAP plots, positive SHAP values indicate a higher likelihood of GDM, whereas negative SHAP values reflect a lower predicted risk.

A. Dataset-Specific Interpretations

The SHAP summary plot (Fig. 3) illustrates the global feature contributions to GDM risk predictions across the dataset. It offers a compact view of how feature values shift the prediction toward or away from GDM for all patients. The y-axis lists features, while the x-axis shows SHAP values, which represent a feature's impact on the model output for each case: positive values push the prediction toward GDM, negative values toward non-GDM, and values farther from zero indicate stronger effects. Each dot corresponds to one patient's contribution for that feature. Dot color encodes the raw feature value for that patient, with blue indicating low values and red indicating high values as shown in the legend. When red points cluster on the positive side, higher feature values are associated with higher predicted risk; when red points cluster on the negative side, higher values are associated with lower predicted risk; mixed colors on both sides suggest non-linear or context-dependent effects.

From a general health perspective, higher BMI continues to exhibit a strong and consistent positive SHAP trend, confirming that elevated adiposity contributes directly to higher GDM risk. In contrast, maternal age presents a less uniform pattern: while younger age values tend to cluster around zero and negative SHAP values (indicating lower predicted risk), older age values are associated with a wider SHAP range, suggesting that age may either increase or decrease risk depending on the presence of co-occurring factors such as BMI or blood pressure. Higher Hemoglobin levels are associated with increased predicted GDM risk, whereas moderate and lower levels generally correspond to lower SHAP values, indicating reduced model-predicted risk.

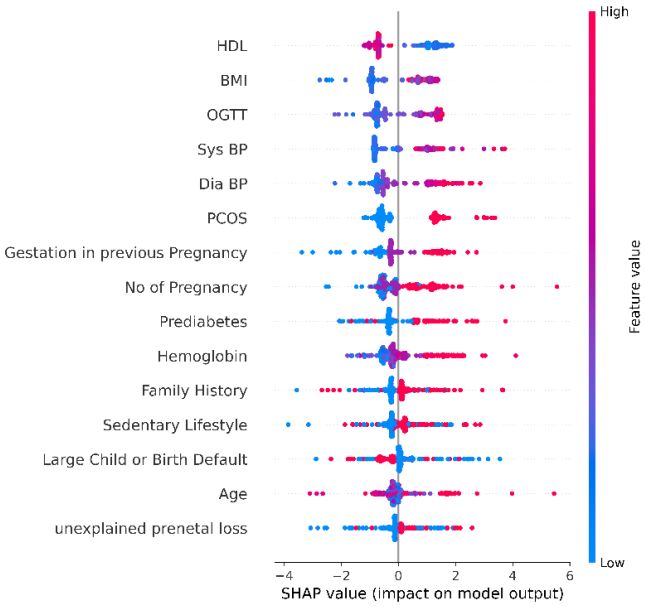


Fig. 3. SHAP summary plot.

For cardiovascular status, both Sys BP and Dia BP display positive SHAP values at elevated readings, indicating that hypertensive disorders in pregnancy often co-occur with insulin resistance and elevated GDM risk. From a metabolic and endocrine perspective, elevated OGTT values substantially increase the predicted risk of GDM. Similarly, the presence of PCOS contributes strong positive SHAP values, indicating higher risk. In contrast, HDL shows a clear inverse association, where lower levels correspond to an increased likelihood of GDM.

Features related to obstetric and reproductive history include Gestation in Previous Pregnancy and Number of Pregnancies, both of which tend to show positive contributions at higher values. A positive history of Unexplained Prenatal Loss is also associated with increased predicted risk. In contrast, Large Child or Birth Default exhibits a more ambiguous pattern, with both its presence and absence producing mixed SHAP contributions. This suggests a feature interaction, where the model may shift predictive importance to other highly correlated variables, such as BMI, maternal age, or parity, thereby diminishing the apparent effect of this obstetric factor.

Finally, Prediabetes, Family History and Sedentary Lifestyle show SHAP patterns where their presence typically contributes to a higher predicted risk of GDM. However, in a small number of cases, even when these risk factors are present, they are linked to lower SHAP values (lower GDM risk), likely due to the dominant influence of other protective factors or interactions that offset their individual impact.

B. Factor Interactions and Dataset-Specific Thresholds

This subsection highlights how clinical and obstetric factors interact to shape the model's GDM risk predictions. SHAP dependence plots (Fig. 4 to Fig. 18) show how the effect of one feature varies with another's value and surface the strongest interactions, helping to spot practical thresholds and non-linear links useful for clinical review. In each plot, the x-axis is the primary feature value, the y-axis is its SHAP value (its contribution to predicted risk), each point is a single case, and a blue-to-red color scale encodes the value of a second interacting feature. Read together, these plots clarify combined patterns such as high BMI with low HDL, or multiple pregnancies with elevated blood pressure, that contribute to higher predicted risk and are expressed in familiar clinical terms to support decision-making.

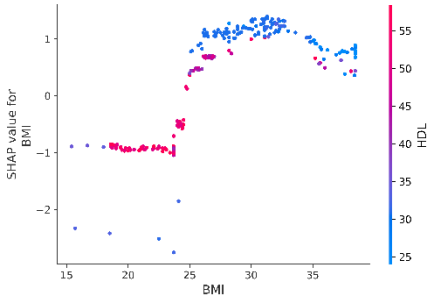


Fig. 4. SHAP dependence plot (BMI & HDL).

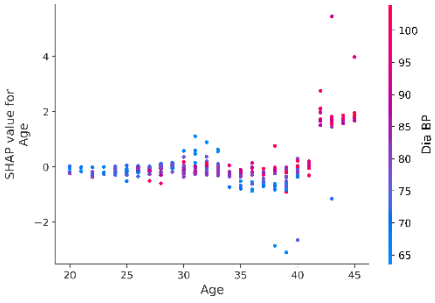


Fig. 5. SHAP dependence plot (Age & Dia BP).

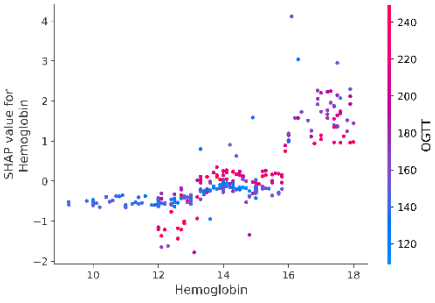


Fig. 6. SHAP dependence plot (Hemoglobin & OGTT).

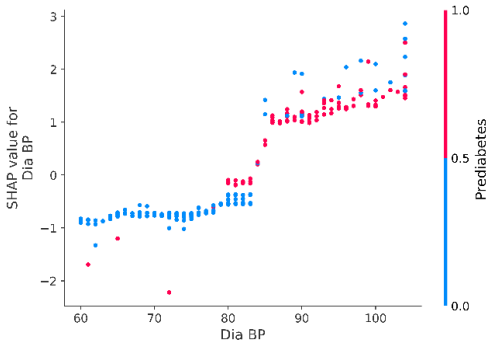


Fig. 7. SHAP dependence plot (Dia BP & Prediabetes).

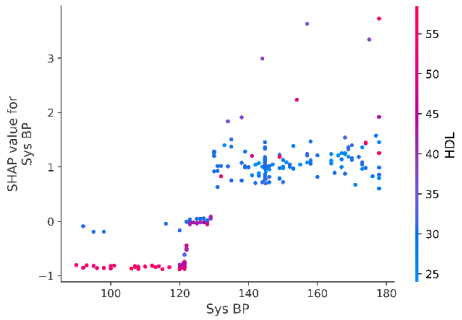


Fig. 8. SHAP dependence plot (Sys BP & HDL).

In Fig. 4, BMI shows a clear threshold around 25. Below this value, SHAP contributions are typically negative, often accompanied by high HDL levels, indicating lower risk. Beyond 25, SHAP values rise sharply, especially in individuals with low HDL. For Maternal Age (Fig. 5), younger women tend to have neutral or slightly negative SHAP values, whereas women above 40 years consistently show positive contributions and are associated with elevated Dia BP. In Fig. 6, Hemoglobin contributes positively to GDM risk between 13 and 16 g/dL, but only when OGTT values are elevated, whereas lower OGTT levels are linked to reduced risk. At levels above 16 g/dL, Hemoglobin consistently shows positive contributions to GDM risk and is associated with high OGTT values. This pattern is important and may be explained by physiological changes such as increased blood viscosity [31], indicating that both factors are likely working together rather than high Hemoglobin level alone.

In the cardiovascular domain, both Dia BP and Sys BP exhibit clear threshold behaviour. Fig. 7 illustrates that Dia BP values above 85 mmHg are associated with positive SHAP contributions, particularly in individuals with prediabetes, whereas values below 80 mmHg in the absence of prediabetes indicate lower risk. Similarly, Fig. 8 shows that Sys BP contributes to higher risk beyond 130 mmHg, with stronger effects observed in patients with low HDL. In contrast, values below 120 mmHg combined with high HDL levels are generally linked to lower risk.

Among metabolic and endocrine features, Fig. 9 shows that OGTT values rise sharply in SHAP contributions beyond 160 mg/dL, particularly in patients with high Sys BP, whereas values below 150 mg/dL are considered protective. In Fig. 10, PCOS consistently contributes to higher predicted risk when present, with its effect being stronger in individuals with high BMI. Fig. 11 illustrates that HDL values below 35 mg/dL are associated with increased risk, especially in women with prior gestations, while values above 35 mg/dL correspond to protective SHAP contributions. This pattern suggests that HDL remains a reliable marker of metabolic health, regardless of gestational history.

Obstetric and reproductive history features show interpretable patterns. Fig. 12 illustrates that Gestation in Previous Pregnancy contributes positively when equal to or greater than two, particularly in patients with elevated Dia BP. Fig. 13 shows that Number of Pregnancies follows a similar upward trend, with risk increasing beyond three. Fig. 14 indicates that Unexplained Prenatal Loss, when present, raises the probability of GDM, although its absence does not necessarily imply low risk. Fig. 15 demonstrates that Prediabetes, when combined with high BMI, is linked to higher GDM risk, whereas its absence together with low BMI is associated with reduced risk.

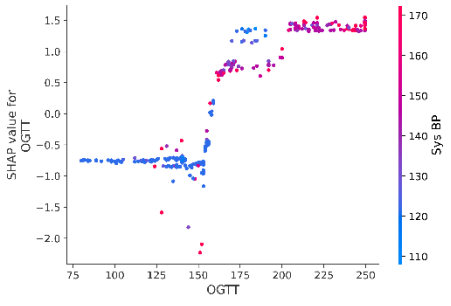


Fig. 9. SHAP dependence plot (OGTT & Sys BP).

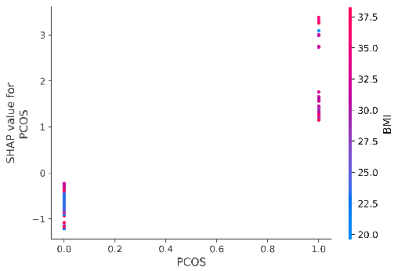


Fig. 10. SHAP dependence plot (PCOS & BMI).

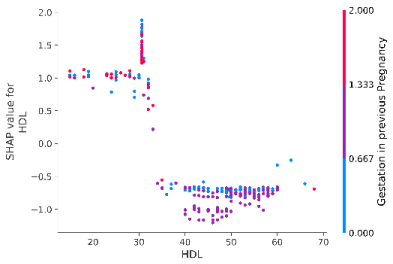


Fig. 11. SHAP dependence plot (HDL & gestation in previous pregnancy).

In contrast, Fig. 16 to Fig. 18 (Large Child or Birth Default, Family History, and Sedentary Lifestyle) show less consistent behavior. Their SHAP contributions vary widely, suggesting that their predictive influence may be partly absorbed by stronger correlated factors.

Collectively, these dependence plots reveal that individual factors influence risk through threshold effects and interaction with other variables. They reinforce clinical understanding of GDM risk while offering finer-grained, data-driven insight into how multiple features combine to shape the model's predictions. Table V summarizes these insights by aligning the data-derived thresholds from SHAP analysis with established clinical reference ranges, highlighting key feature interactions and their associated GDM risk. This comparative view helps contextualize model behavior against known medical standards, reinforcing the interpretability and clinical relevance of EYE-GDM results.

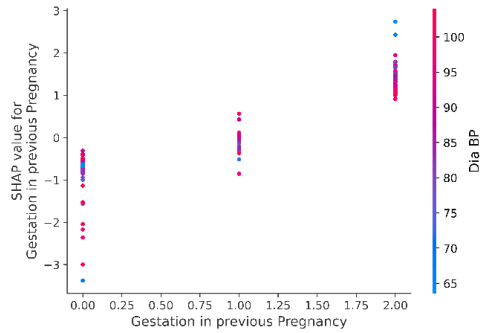


Fig. 12. SHAP dependence plot (gestation in previous pregnancy & Dia BP).

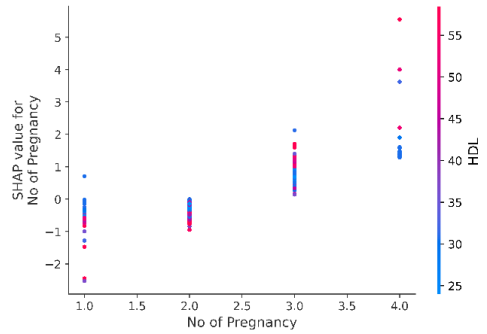


Fig. 13. SHAP dependence plot (number of pregnancy & HDL).

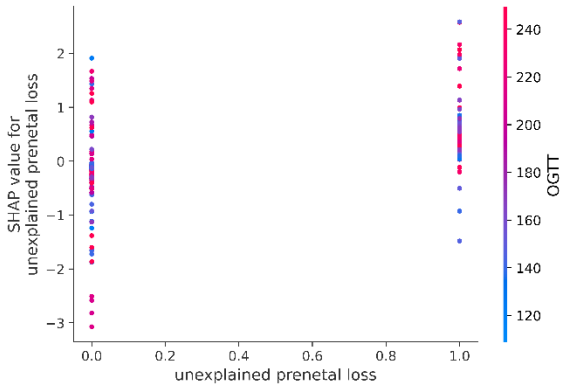


Fig. 14. SHAP dependence plot (unexplained prenatal loss & OGTT).

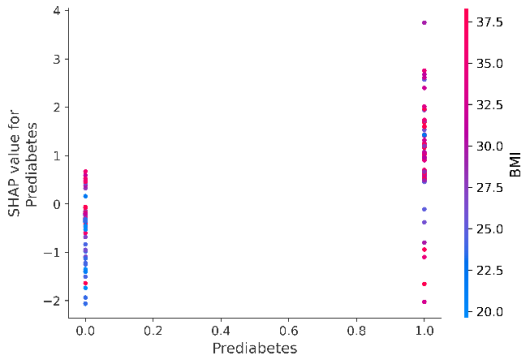


Fig. 15. SHAP dependence plot (prediabetes & BMI).

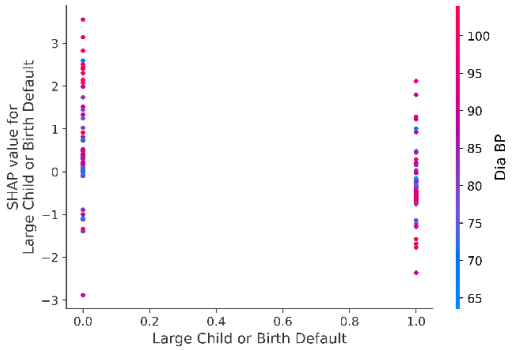


Fig. 16. SHAP dependence plot (large child or birth default & Dia BP).

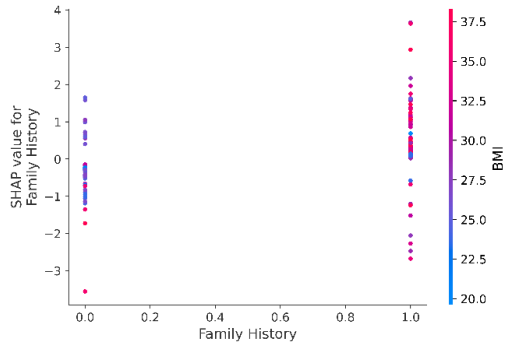


Fig. 17. SHAP dependence plot (family history & BMI).

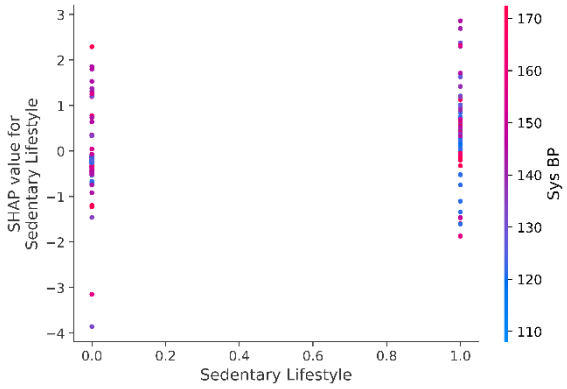


Fig. 18. SHAP dependence plot (sedentary lifestyle & sys BP).

TABLE V. DATASET-DERIVED THRESHOLDS, INTERACTING FACTORS, AND CLINICAL NORMAL RANGES FOR KEY GDM PREDICTORS

Feature	Clinical Cutoff / Normal Range	Data-driven Threshold	Interacts with	GDM Risk
BMI	18.5–24.9 kg/m² [32], [33]	> 25	Low HDL	High
		< 25	High HDL	Low
Age	-	> 40 years	High Dia BP	High
Hemoglobin	12-16 g/dL [34]	13-16 g/dL	High OGTT	
		> 16 g/dL	-	
Dia BP	< 80 mmHg (normal), 80-89 mmHg (Stage 1 hypertension), ≥ 90 mmHg (Stage 2 hypertension) [35]	> 85 mmHg	Prediabetes Presence	
		< 80 mmHg	Prediabetes Absence	Low
Sys BP	< 120 mmHg (normal), 120-129 mmHg (elevated), 130-139 mmHg (Stage 1 hypertension), ≥ 140 mmHg (Stage 2 hypertension) [35]	> 130 mmHg	Low HDL	High
		< 120 mmHg	High HDL	Low
OGTT	75-g 2-h: <153 mg/dL [32]	> 160 mg/dL	High Sys BP	High
		< 150 mg/dL	-	Low
PCOS	-	Presence	High BMI	High
HDL	50-80 mg/dL (women) [36]	< 35 mg/dL	Gestational History	
		> 35 mg/dL	-	Low
Gestation in Previous Pregnancy	-	≥ 2	High Dia BP	High
No of Pregnancies	-	≥ 3	-	
Unexplained Prenatal Loss	-	Presence	-	
Prediabetes	-	Presence	High BMI	

C. Patient-Specific Interpretations

To complement the global insights, we examine the model's local reasoning with SHAP waterfall plots. A waterfall plot (Fig. 19) decomposes one patient's prediction by starting at the baseline E[f(x)] (the model's expected output over the training background) and then applying each feature's SHAP contribution in sequence until it reaches the patient's final prediction f(x). The x-axis represents the model output scale, so movement to the right increases the prediction and movement to the left decreases it; the y-axis lists the patient's features with their observed values. Bars show the size and direction of each contribution, with red increasing risk and blue decreasing risk.

For this patient, a 24-year-old woman with BMI 23.69 and no sedentary lifestyle, PCOS, prediabetes, or family history (see Table VI), the predicted probability of GDM is f(x) = 0.001, well below the dataset baseline E[f(x)] = 0.612, indicating a low-risk profile. The model arrives at this result through several negative contributions: the absence of unexplained prenatal loss (−0.09), BMI below 25, systolic blood pressure 117 mmHg (below 130 mmHg), and OGTT 140.32 mg/dL (below clinical and data-driven thresholds). Additional decreases come from HDL 53 mg/dL (above threshold) and diastolic blood pressure 80 mmHg (below 85 mmHg).

Her parity of two is below the higher-risk cutoff of three, and the absence of PCOS further reduces risk. Although hemoglobin is 9.8 g/dL, which is below the range where positive SHAP effects were observed in this dataset (>16 g/dL), it does not raise predicted GDM risk here; however, it remains clinically low and may warrant separate attention during pregnancy.

This case is consistent with the data-driven thresholds that we presented earlier in Table V: none of her values exceed the risk cutoffs, and several interacting protective features (such as high HDL, normal blood pressure, and OGTT below threshold) contribute cumulatively to a low-risk prediction. This case highlights how the model synthesizes individual-level clinical factors to make well-reasoned predictions. It also demonstrates SHAP's utility in explaining not only high-risk cases but also in identifying low-risk profiles by tracing the impact of protective factors.

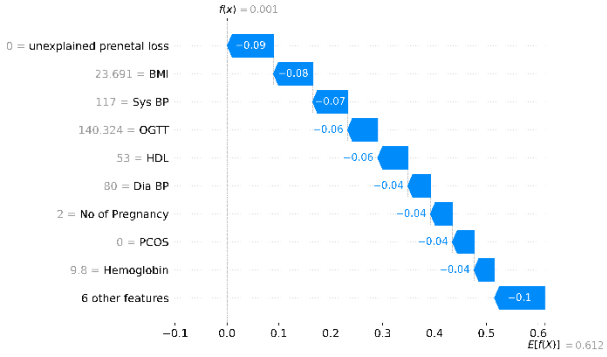


Fig. 19. SHAP waterfall plot for a representative patient.

TABLE VI. PATIENT-LEVEL FEATURE VALUES FOR LOCAL EXPLANATION

Feature	Value
Age	24
No. of Pregnancy	2
Gestation in previous pregnancy	None
BMI	23.96
HDL	53
Family History	None
Unexplained prenatal loss	None
Large Child or Birth Default	None
PCOS	None
Sys BP	117
Dia BP	80
OGTT	140.32
Hemoglobin	9.8
Sedentary Lifestyle	None
Prediabetes	None

VI. Discussion

This work demonstrates how interpretable ensemble modelling, implemented via the EYE-GDM system, can support risk prediction in the context of GDM. We specifically selected GDM as the application focus for several reasons. First, GDM presents a clinically significant yet temporally constrained condition, with substantial risks for both mother and infant, including preeclampsia, macrosomia, neonatal hypoglycaemia, and long-term progression to type 2 diabetes [4]. Its transient onset during pregnancy necessitates timely and explainable risk assessments that clinicians can trust.

From a scientific perspective, the contributions of this work lie in the methodological refinements introduced in EYE-GDM, including the reconfigured ensemble structure, the explanation-weighted meta-learning approach, and the domain-aligned interpretability evaluation. These represent technical extensions of the EYE framework. The applied contribution lies in demonstrating these methodological advances within the maternal health domain, where EYE-GDM provides clinically interpretable risk assessments, data-driven thresholds, and patient-level insights specific to gestational diabetes.

EYE-GDM supports interpretation at both the dataset and patient levels. At the dataset level, SHAP summary plots and dependence plots allow for the identification of global risk patterns, data-driven thresholds, and feature interactions. These insights capture not only the individual impact of clinical features but also how their interactions influence GDM risk. Such results can inform strategic clinical guidelines, such as refining screening protocols for high-risk groups. In contrast, patient-level explanations, visualized through SHAP waterfall plots, help clinicians understand why a specific patient is predicted to be at risk, thereby supporting individualized intervention planning. Together, these interpretation levels enhance the model's clinical utility and support population-level decision-making and personalized care.

The predictive performance of EYE-GDM is competitive, achieving an accuracy of 0.9789, an F1-score of 0.9790, a precision of 0.9747, a recall of 0.9833, and an AUC-ROC of 0.9981. These results outperform previously reported models on the same dataset while simultaneously offering transparent explanations. Compared with conventional ensemble or deep-learning models used for GDM prediction, EYE-GDM provides distinct advantages by integrating interpretability directly within the training process rather than applying it as a separate post-hoc step. This ensures that explanations remain consistent with the ensemble's internal reasoning. In addition, the model combines global and patient-level reasoning within a single framework, enabling both population insights and individual clinical interpretation, which is rarely achieved by previous methods. Furthermore, incorporating expert clinical feedback during validation strengthens the model's credibility and aligns its explanations with real-world diagnostic reasoning and clinical practice.

More importantly, the interpretations derived from SHAP analysis are well-aligned with established medical knowledge, confirming the validity of the model's outputs. SHAP values for Age increase after 40, consistent with the clinical evidence that advanced maternal age is a GDM risk factor [37], [38]. Risk rises significantly for BMI values above 25, aligning with WHO thresholds for overweight [32], [33]. OGTT values show a sharp SHAP increase beyond 160 mg/dL, supporting diagnostic thresholds for impaired glucose tolerance [32]. Hemoglobin contributes to risk in the 13–16 g/dL range only when OGTT is also elevated, underscoring the need to interpret it contextually. HDL levels above 35 mg/dL exhibit protective SHAP values, matching its known cardiometabolic role [36]. Similarly, Sys BP starts contributing to higher risk beyond 130 mmHg, with stronger effects in patients who have low HDL. This pattern reflects the well-known clustering of high blood pressure, abnormal lipids, and elevated glucose seen in metabolic syndrome [39]. Prediabetes and PCOS both contribute positively to GDM risk, consistent with medical literature [40], [41]. Finally, obstetric history factors such as prior gestation with complications and a higher number of pregnancies display threshold-based patterns, reinforcing their importance in GDM prognosis [42], [43].

The interpretability results revealed clinically consistent yet subtle patterns. Interactions such as BMI with HDL and blood pressure with metabolic indicators showed that GDM risk emerges from the combined effects of metabolic and cardiovascular factors rather than single predictors. The model also identified context-dependent relationships, where certain features influence risk only in combination with others, reflecting the complexity of gestational physiology. The close alignment between model-derived thresholds (BMI 25 kg/m², Dia BP 85 mmHg, Sys BP 130 mmHg) and established reference ranges supports the clinical credibility of EYE-GDM outputs. These findings illustrate how interpretable AI can mirror clinical reasoning in maternal health, providing insights that clarify how metabolic, cardiovascular, and obstetric factors interact to shape individual risk. In practice, the identified thresholds can inform more targeted antenatal screening and help refine early intervention strategies for high-risk patients. The variability observed in lifestyle and family-history features highlights the need for broader, standardized datasets to enhance robustness and ensure the generalizability of model insights across populations.

Clinician validation further reinforced these interpretations. The clinicians confirmed that the model's predictions were consistent with established GDM risk factors and reflected the multifactorial nature of clinical practice, where risk factors interact rather than act independently. They noted, however, that hemoglobin alone is not an independent risk factor for GDM, whereas HbA1c is a more reliable predictor. The observed association between elevated hemoglobin levels and higher OGTT values in this dataset was considered important, potentially reflecting physiological mechanisms such as increased blood viscosity [31], suggesting that these factors act in combination rather than through hemoglobin alone. The clinicians also emphasized the need for clearer dataset documentation and recommended involving clinical experts in the data collection process to ensure the inclusion of key GDM biomarkers, such as HbA1c and different OGTT test types. The validation was regarded as valuable both for strengthening trust and for guiding iterative refinement.

While the model showed strong alignment with clinical reasoning, certain limitations remain. This work extends prior

validation of the EYE framework (EYE-WD) to a distinct clinical population, demonstrating its adaptability across domains. However, because the dataset represents a single regional population, sociocultural and healthcare-access factors that influence screening and diagnosis may have affected model behavior. Future work will focus on validating the model across multi-centre and demographically diverse GDM cohorts as such datasets become available.

The results of this work highlight the importance of embedding interpretability directly into model design for maternal healthcare. EYE-GDM demonstrates how interpretable models can uncover clinically relevant patterns that align with established medical knowledge, advancing maternal health AI by combining high predictive accuracy with clinically grounded explanations.

VII. Conclusion and Future Work

EYE-GDM, an implementation of the EYE framework for gestational diabetes risk prediction, delivers strong predictive performance together with clinically meaningful and interpretable outputs. It directly addresses two long-standing gaps in prior work: the tendency to prioritize predictive accuracy without producing explanations that clinicians can apply, and the reliance on post-hoc interpretability methods that fail to capture how models reason internally. By embedding SHAP explanations into the training pipeline, EYE-GDM ensures that each base learner is explained individually and that their contributions are integrated through logistic regression weights, producing ensemble-level reasoning consistent with the model's internal decision logic.

EYE-GDM achieved competitive results, surpassing earlier reports on the same dataset. More importantly, its interpretable outputs were consistent with established medical knowledge, showing that the framework can provide reliable risk prediction while also generating explanations of practical value in maternal care. This combination strengthens confidence in the system as a tool that clinicians can trust.

Beyond its technical contributions, EYE-GDM advances the field of explainable healthcare AI by demonstrating how interpretability can be embedded throughout the model development process rather than appended after training. This integration shifts explainability from a diagnostic tool to a design principle, enabling models that reason transparently while maintaining predictive strength. Within maternal health, EYE-GDM illustrates how such integration can transform complex risk prediction into a clinically meaningful, trust-enhancing process, setting a foundation for future interpretable systems across other sensitive domains.

Future work will continue engagement with clinical experts to refine interpretability outputs and ensure their clarity in practice. Additional directions include extending the framework to other healthcare conditions where interpretability is essential, optimizing it for real-time deployment, and evaluating its performance across diverse populations.

The EYE framework builds on our earlier FIXAIH roadmap, which set out six design priorities for developing trustworthy and interpretable AI in healthcare. While FIXAIH defines high-level goals, EYE operationalizes selected elements into a structured and testable framework centered on interpretability by design, transparent development, and clinical validation. EYE-GDM demonstrates how these principles can be applied in maternal health, while EYE-WD shows their application in women's diabetes. Together, these implementations highlight how domain-specific frameworks can serve as practical steps toward the broader FIXAIH vision, linking conceptual guidance with deployment-ready systems that advance clinical care while maintaining transparency and trust.

Acknowledgment

This study is derived from a research grant funded by the Research, Development, and Innovation Authority (RDIA), Kingdom of Saudi Arabia, with grant number 12615-iu-2023-IU-R-2-1-EI.

References

[1] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," Nat Med, vol. 28, no. 1, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.

[2] P. Kumar, S. Chauhan, and L. K. Awasthi, "Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions," Eng Appl Artif Intell, vol. 120, p. 105894, Apr. 2023, doi: 10.1016/J.ENGAPPAI.2023.105894.

[3] Y. Du, A. M. Antoniadi, C. McNestry, F. M. McAuliffe, and C. Mooney, "The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations," Applied Sciences, vol. 12, no. 20, 2022, doi: 10.3390/app122010323.

[4] World Health Organisation (WHO), "Diabetes," WHO. Accessed: Mar. 01, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[5] S. Feduniw, D. Sys, S. Kwiatkowski, and A. Kajdy, "Application of artificial intelligence in screening for adverse perinatal outcomes: A protocol for systematic review," Medicine, vol. 99, no. 50, p. e23681, Dec. 2020, doi: 10.1097/MD.0000000000023681.

[6] K. S. Lee and K. H. Ahn, "Application of artificial intelligence in early diagnosis of spontaneous preterm labor and birth," Diagnostics, vol. 10, no. 9, p. 733, 2020, doi: 10.3390/diagnostics10090733.

[7] M. Becker et al., "Revealing the impact of lifestyle stressors on the risk of adverse pregnancy outcomes with multitask machine learning," Front Pediatr, vol. 10, p. 933266, 2022, doi: 10.3389/fped.2022.933266.

[8] H. Y. Lu et al., "Digital Health and Machine Learning Technologies for Blood Glucose Monitoring and Management of Gestational Diabetes," IEEE Rev Biomed Eng, vol. 17, pp. 98–117, 2024, doi: 10.1109/RBME.2023.3242261.

[9] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif Intell, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[10] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit Health, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[11] S. Alghamdi, R. Mehmood, F. Alqurashi, T. Alghamdi, A. AlAhmadi, and S. Ghazali, "EYE and EYE-WD: Clinically Validated, Interpretable Ensemble Learning for Women's Diabetes," SSRN Preprint, Aug. 2025, doi: 10.2139/ssrn.5386582.

[12] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," in Proceedings of the 4th Machine Learning for Healthcare Conference, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., in Proceedings of Machine Learning Research, vol. 106. PMLR, Sep. 2019, pp. 359–380. [Online]. Available: https://proceedings.mlr.press/v106/tonekaboni19a.html

[13] A. Sumathi, S. Meganathan, and B. V. Ravisankar, "An Intelligent Gestational Diabetes Diagnosis Model Using Deep Stacked Autoencoder.," Computers, Materials & Continua, vol. 69, no. 3, pp. 3109–3126, Mar. 2021, doi: 10.32604/cmc.2021.017612.

[14] A. Sumathi and S. Meganathan, "Ensemble Classifier Technique to Predict Gestational Diabetes Mellitus (GDM).," Computer Systems Science Engineering, vol. 40, no. 1, pp. 313–325, 2022, doi: 10.32604/csse.2022.017484.

[15] R. Jader and S. Aminifar, "Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach," Applied Computational Intelligence and Soft Computing, vol. 2022, no. 1, 2022, doi: 10.1155/2022/9749579.

[16] H. Gallardo-Rincón et al., "MIDO GDM: an innovative artificial intelligence-based prediction model for the development of gestational diabetes in Mexican women," Sci Rep, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-34126-7.

[17] T. Zheng et al., "A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women," BMC Pregnancy Childbirth, vol. 19, no. 1, 2019, doi: 10.1186/s12884-019-2374-8.

[18] J. Shen et al., "An innovative artificial intelligence-based app for the diagnosis of gestational diabetes mellitus (GDM-AI): Development study," J Med Internet Res, vol. 22, no. 9, p. e21573, Sep. 2020, doi: 10.2196/21573.

[19] Y. T. Wu et al., "Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning," Journal of Clinical Endocrinology & Metabolism, vol. 106, no. 3, 2021, doi: 10.1210/clinem/dgaa899.

[20] X. Hu, X. Hu, Y. Yu, and J. Wang, "Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm," Front Endocrinol (Lausanne), vol. 14, 2023, doi: 10.3389/fendo.2023.1105062.

[21] M. Watanabe, A. Eguchi, K. Sakurai, M. Yamamoto, and C. Mori, "Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan Environment and Children's Study," Sci Rep, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-44313-1.

[22] Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," Sci Rep, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-05112-2.

[23] V. Vivek Khanna et al., "Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers," Cogent Eng, vol. 11, no. 1, p. 2330266, Dec. 2024, doi: 10.1080/23311916.2024.2330266.

[24] H. Zaky et al., "Machine learning based model for the early detection of Gestational Diabetes Mellitus," BMC Med Inform Decis Mak, vol. 25, no. 1, p. 130, 2025, doi: 10.1186/s12911-025-02947-3.

[25] S. Alghamdi, R. Mehmood, F. A. Alqurashi, and A. Alzahrani, "Paving the Roadmap for XAI and IML in Healthcare: Data-Driven Discoveries and the FIXAIH Framework," IEEE Access, vol. 13, pp. 174393–174427, 2025, doi: 10.1109/ACCESS.2025.3616353.

[26] A.sumathi and S.Meganathan, "Gestational Diabetes Mellitus (GDM Dataset)." Accessed: Jul. 10, 2025. [Online]. Available: https://www.kaggle.com/dsv/3245285

[27] S. Sperandei, "Understanding logistic regression analysis," Biochem Med (Zagreb), vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. , Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.

[29] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat Mach Intell, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/S42256-019-0138-9.

[30] S. H. Golden, C. Yajnik, S. Phatak, R. L. Hanson, and W. C. Knowler, "Racial/ethnic differences in the burden of type 2 diabetes over the life course: a focus on the USA and India," Diabetologia, vol. 62, no. 10, pp. 1751–1760, 2019, doi: 10.1007/s00125-019-4968-0.

[31] S. Rafaqat and S. Rafaqat, "Role of hematological parameters in pathogenesis of diabetes mellitus: A review of the literature," World Journal Hematology, vol. 10, no. 3, pp. 25–41, Mar. 2023, doi: 10.5315/WJH.V10.I3.25.

[32] American Diabetes Association Professional Practice Committee, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025," Diabetes Care, vol. 48, no. Supplement_1, pp. S27–S49, Jan. 2025, doi: 10.2337/dc25-S002.

[33] World Health Organization, "Obesity and overweight," May 2025. Accessed: Sep. 18, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[34] National Board of Medical Examiners, "Laboratory Reference Values," Mar. 2025. Accessed: Oct. 18, 2025. [Online]. Available: https://www.nbme.org/laboratory-values

[35] P. K. Whelton et al., "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines," J Am Coll Cardiol, vol. 71, no. 19, pp. 2199–2269, May 2018, doi: 10.1016/j.jacc.2017.11.005.

[36] Cleveland Clinic, "HDL: Why It's 'Good' Cholesterol." Accessed: Sep. 10, 2025. [Online]. Available: https://my.clevelandclinic.org/health/articles/24395-hdl-cholesterol

[37] A. Bouzaglou et al., "Pregnancy at 40 years Old and Above: Obstetrical, Fetal, and Neonatal Outcomes. Is Age an Independent Risk Factor for Those Complications?," Front Med (Lausanne), vol. 7, p. 208, May 2020, doi: 10.3389/fmed.2020.00208.

[38] E. A. AlJahdali and N. S. AlSinani, "Pregnancy outcomes at advanced maternal age in a tertiary Hospital, Jeddah, Saudi Arabia," Saudi Med J, vol. 43, no. 5, p. 491, May 2022, doi: 10.15537/smj.2022.43.5.20220023.

[39] J. Kim, "Metabotype Risk Clustering Based on Metabolic Disease Biomarkers and Its Association with Metabolic Syndrome in Korean Adults: Findings from the 2016–2023 Korea National Health and Nutrition Examination Survey (KNHANES)," Diseases, vol. 13, no. 8, p. 239, 2025, doi: 10.3390/diseases13080239.

[40] G. Wilkie, E. Delpapa, and H. Leftwich, "Early Diagnosis of Prediabetes among Pregnant Women that Develop Gestational Diabetes Mellitus and Its Influence on Perinatal Outcomes," Am J Perinatol, vol. 41, no. 3, pp. 343–348, 2024, doi: 10.1055/A-1682-2643.

[41] H. J. Kim, E. H. Kim, E. Ko, S. Park, and Y. Lee, "The Impact of Polycystic Ovary Syndrome on Gestational Diabetes Mellitus, Disease Knowledge, and Health Behaviors," Healthcare, vol. 13, no. 7, 2025, doi: 10.3390/healthcare13070717.

[42] J. Pukkila et al., "The recurrence risk of gestational diabetes according to the number of abnormal values in the oral glucose tolerance test," Acta Obstet Gynecol Scand, vol. 104, no. 8, pp. 1452–1462, Aug. 2025, doi: 10.1111/AOGS.15148.

[43] B. Liu et al., "Higher Numbers of Pregnancies Associated With an Increased Prevalence of Gestational Diabetes Mellitus: Results From the Healthy Baby Cohort Study," J Epidemiol, vol. 30, no. 5, pp. 208–212, 2020, doi: 10.2188/jea.JE20180245.