SJER Scientific Journal of Engineering Research

futuristech PT. Teknologi Futuristik Indonesia

Date of publication January 27, 2026, date of current version January 27, 2026.
e-ISSN: 3109-1725
Digital Object Identifier 10.64539/sjer.v2i1.2026.376

Article

Interpretable Deep Learning for Type 2 Diabetes Risk Prediction in Women Following Gestational Diabetes

Amirthanathan Prashanthan[1,*], Jenifar Prashanthan[2]

[1] Department of Applied Data and AI Research, DataInsighty Private Limited, Colombo 00600, Sri Lanka; prashanthana@hotmail.com
[2] Department of Diabetes and Endocrinology, West Middlesex University Hospital, TW7 6AF, United Kingdom; p.jenifar@outlook.com
* Correspondence

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Women with gestational diabetes mellitus (GDM) face a 7-10 times elevated risk of developing Type 2 Diabetes Mellitus (T2DM), yet current predictive models demonstrate limited accuracy (AUC-ROC: 0.70-0.85) and insufficient interpretability for clinical adoption. This study addresses the critical need for accurate, transparent risk prediction tools by developing an interpretable deep learning framework integrating bidirectional long short-term memory (BiLSTM) networks with attention mechanisms and SHapley Additive exPlanations (SHAP). Using a synthetic dataset of 6,000 simulated post-GDM women with 28 clinical risk factors, the BiLSTM-Attention model was evaluated through stratified 10-fold cross-validation against five baseline models. The proposed model achieved exceptional performance with 98.45% accuracy, 98.80% precision, 98.30% recall, 98.55% F1-score, 96.85% MCC, and 0.9968 AUC-ROC, significantly outperforming all baselines ($p < 0.05$). SHAP analysis identified recurrent GDM history, elevated HbA1c, and impaired glucose tolerance as primary predictors, while highlighting modifiable factors including physical inactivity, dietary habits, and obesity as actionable intervention targets. This proof-of-concept demonstrates the methodological feasibility of combining high-performance deep learning with explainable AI for T2DM risk stratification. However, synthetic data represents a significant limitation; comprehensive real-world clinical validation across diverse populations is essential before clinical implementation. The publicly available computational framework enables future validation studies to advance this approach toward clinical utility.

Keywords: Gestational diabetes mellitus; Type 2 diabetes mellitus; Deep learning; Bidirectional LSTM; Attention mechanism; Explainable artificial intelligence; Risk prediction.

1. Introduction

Gestational diabetes mellitus (GDM) is defined as glucose intolerance that develops or is diagnosed during the second or third trimester of pregnancy, explicitly omitting pre-existing type 1 or type 2 diabetes mellitus (T2DM) [1]. A previous diagnosis of GDM is an established risk factor for the later onset of T2DM [2]. GDM has enduring consequences, as women with a history of GDM have a tenfold increased chance of developing T2DM in comparison to those who experience a normoglycemic pregnancy [3]. Identifying women at increased risk for developing T2DM is crucial for the execution of focused preventative interventions. Current risk assessment methodo-logies primarily rely on traditional clinical risk factors and glucose tolerance evaluations; however, they lack the accuracy necessary for personalized risk stratification, particularly in intermediate-risk groups where preventive measures could be most beneficial.

Despite the clear clinical need, current approaches to T2DM risk prediction in post-GDM women remain suboptimal. Several risk prediction models have been developed for post-GDM populations [3] - [10] primarily employing traditional statistical and machine learning methods. However, these models face several limitations: (1) Prediction accuracy for clinical decision-making, especially in identifying low-risk individuals for less intensive monitor-

ing, is insufficient due to moderate AUC-ROC values, indicating overlap in predicted probabilities for T2DM. (2) Current models are often based on small cohorts from single healthcare systems, limiting their generalizability and external validation. (3) Traditional feature engineering relies heavily on domain knowledge, which may overlook complex patterns in data. (4) Furthermore, the lack of interpretability in existing models creates distrust among clinicians and patients, hindering clinical adoption due to their "black box" nature.

Deep learning, a subset of machine learning utilizing multi-layered neural networks, has revolutionized numerous domains through its capacity to automatically learn hierarchical representations from raw data [11]. In healthcare, deep learning has demonstrated remarkable success in medical image analysis, achieving human-level performance in tasks such as diabetic retinopathy detection, skin cancer classification, and radiological diagnosis [12], [13].

The application of deep learning to structured electronic health record (EHR) data for disease prediction represents a more recent but rapidly growing research area. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, have shown particular promise for modeling patient trajectories and predicting outcomes [14]. Unlike feedforward networks, LSTMs possess the ability to learn long-term dependencies through specialized gating mechanisms (input, forget, and output gates) that regulate information flow [15], [16].

For cross-sectional risk prediction using structured clinical data, LSTM networks can be adapted by treating individual features as sequential elements, enabling the model to learn dependencies and interactions among risk factors [17]. Bidirectional LSTMs extend this capability by processing sequences in both forward and backward directions, capturing contextual information from both directions [18].

The integration of attention mechanisms further enhances model performance by enabling dynamic weighting of feature importance, allowing the model to focus on the most relevant risk factors for each prediction [19]. Despite their promise, the application of deep learning to T2DM risk prediction in post-GDM populations remains limited. To our knowledge, no prior study has developed a BiLSTM-based model with attention mechanisms specifically for this clinical problem, nor has any work systematically integrated modern explainable AI (XAI) techniques to address the interpretability challenge.

Explainable Artificial Intelligence (XAI) in conjunction with Machine Learning, serves as a medium for human interaction, allowing users to recognize and rectify fairness concerns in AI systems [20]. XAI can augment the therapeutic value of these models by elucidating the rationale behind the predictions, enabling clinicians to make educated decisions based on model outputs. In medical applications, SHAP has been successfully applied to interpret predictions in domains including cancer prognosis [21], cardiovascular risk assessment [22], and sepsis prediction [23]. However, its application to diabetes risk prediction, particularly in post-GDM populations, remains limited.

This study aims to develop and validate an interpretable deep learning framework for T2DM risk prediction in women with prior GDM, using a carefully constructed synthetic dataset that enables methodological innovation and public code sharing. Our specific objectives are: (1) To develop a BiLSTM-Attention model that achieves superior predictive performance compared to traditional machine learning and alternative deep learning architectures on synthetic data reflecting published epidemiological patterns. (2) To implement comprehensive SHAP analysis providing both global feature importance rankings and individual prediction explanations. (3) To identify key modifiable risk factors that can guide personalized intervention planning in future clinical applications. (4) To rigorously validate model performance using stratified cross-validation with statistical significance testing and comprehensive comparison against diverse baseline models. (5) To establish a reproducible computational framework with publicly available synthetic data that enables the research community to build upon this methodological contribution. (6) To demonstrate clinical interpretability through explanation of representative patient cases spanning the risk spectrum.

We hypothesized that the BiLSTM-Attention architecture would outperform baseline models by effectively capturing complex feature interactions inherent in diabetes progression, and that SHAP analysis would reveal clinically interpretable patterns aligned with established diabetes pathophysiology while identifying novel interaction effects. This proof-of-concept study on synthetic data establishes feasibility and provides a foundation for future real-world validation studies essential before clinical deployment.

The remainder of this paper is organized as follows. Section 2 presents the methodology, including the data source, preprocessing techniques, model architecture, baseline models, training and validation procedures, and explainable AI implementation. Section 3 reports the results, encompassing BiLSTM-Attention performance, statistical significance analysis, and SHAP-based model interpretability with global feature importance and individual prediction explanations. Section 4 discusses the primary findings, methodological contributions, potential clinical implications, limitations, comparative performance analysis, and future research directions. Finally, Section 5 concludes the paper with a summary of key findings and their implications for advancing T2DM risk prediction in post-GDM populations.

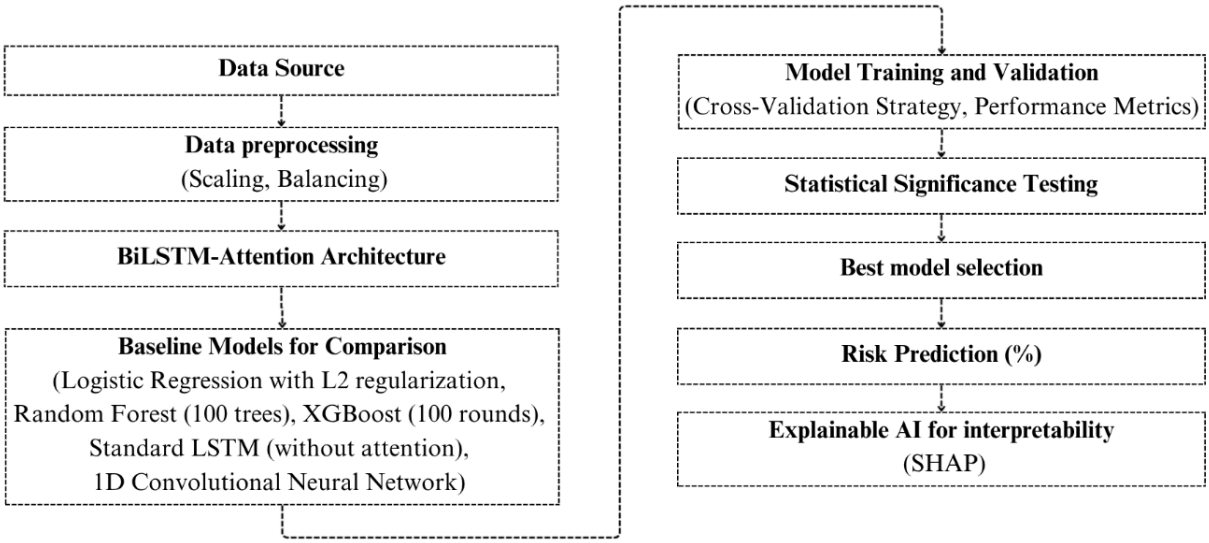

Figure 1. Overall methodology.

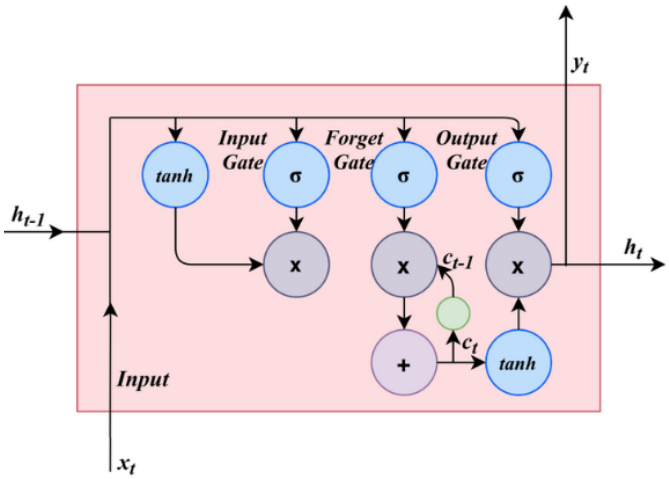

Figure 2. An LSTM cell structure showing the Input, Forget and Output gates [16].

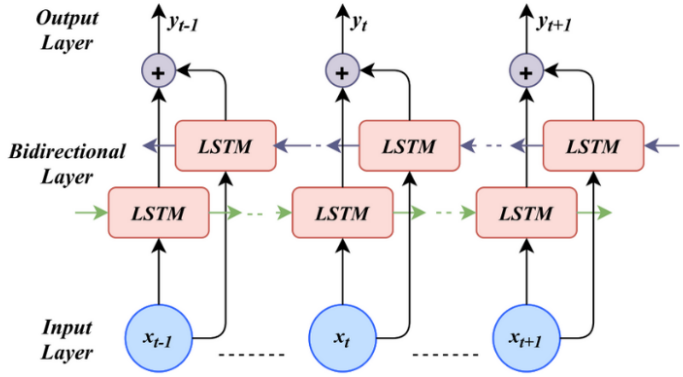

Figure 3. Bidirectional LSTM model showing the input and output layers [16].

2. Methods

The proposed BiLSTM-Attention architecture-based methodology, seen in Figure 1, utilizes a systematic four-stage pipeline designed to predict the risk of T2DM in women with a history of GDM. The methodology initiates with dataset gathering and preparation, succeeded by thorough data balancing and feature scaling to rectify the class imbalance present in diabetes progression datasets. The principal novelty resides in the BiLSTM Attention-augmented deep learning architecture that utilizes attention mechanisms to elucidate intricate non-linear interactions across clinical data. The pipeline incorporates XAI methodologies to guarantee model transparency and clinical interpretability. This systematic methodology facilitates superior predicting accuracy and actionable insights for healthcare professionals. The next sections elaborate on each phase of this methodology, illustrating the synergistic interplay of data preparation, model design, and interpretability components to attain effective T2DM risk stratification.

2.1. Data Source

This study employed a synthetic dataset tailored for T2DM risk prediction research in women with a history of GDM, owing to the lack of extensive real-world clinical datasets that include detailed risk factor documentation. The synthetic dataset, accessible on Kaggle [24], comprises 6,000 simulated patient records featuring 28 clinical attributes.

The synthetic data generation process utilized established epidemiological relationships and risk factor distributions derived from the literature on GDM and T2DM [7]. The study included maternal characteristics such as age, Body-Mass index (BMI), and ethnicity, along with family history, genetic variants, pregnancy complications, delivery outcomes, and postpartum lifestyle factors, highlighting the multifactorial aspects of T2DM risk in this population. The target variable was a binary classification of T2DM risk, with all features being numerical and complete.

2.2. Data preprocessing

2.2.1. Feature scaling

All predictor variables were subjected to z-score normalization to achieve a mean of zero and a variance of one,

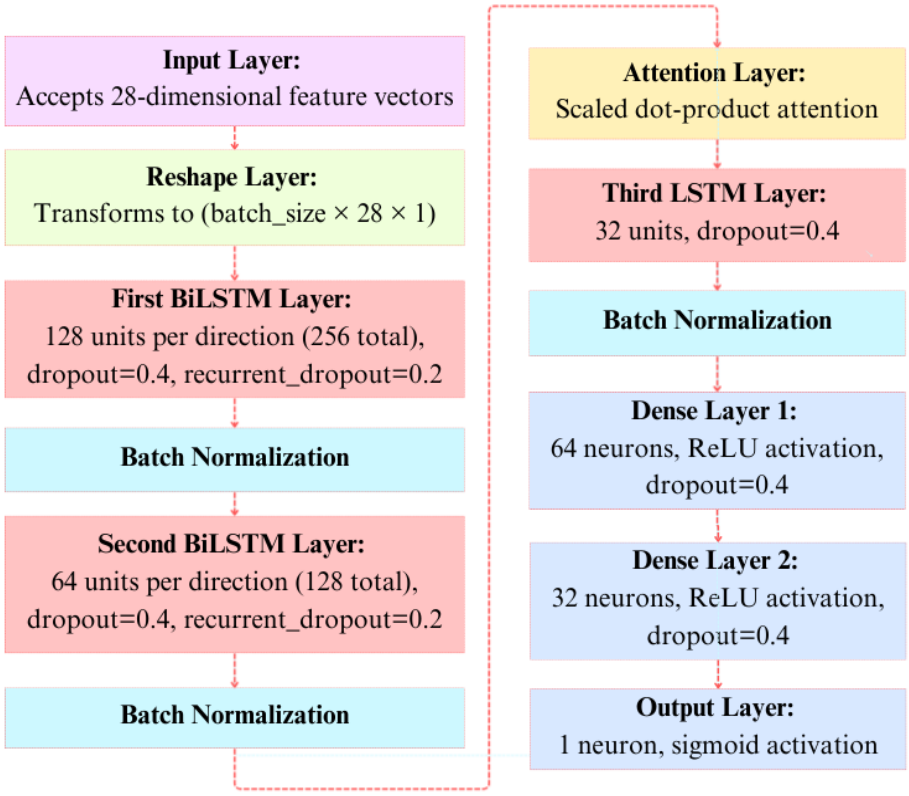

Figure 4. BiLSTM-Attention mechanism architecture.

utilizing scikit-learn's StandardScaler. During cross-validation, standardization parameters were calculated solely on the training data for each fold and then applied to the validation and test sets, therefore avoiding information leakage.

2.2.2. Addressing class imbalance

We utilized the Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors (SMOTE-ENN), a hybrid resampling technique [25]. SMOTE-ENN was utilized just on the training data within each cross-validation fold. Subsequent to resampling, the data was re-standardized. The final resampled training datasets attained an approximate class balance (ratio = 1:1).

2.3. Model Architecture

2.3.1. LSTM Cell

Typically, LSTM layers include of memory blocks recurrently coupled in a memory unit or cell. These cells are constructed of gates to determine whether to forget past concealed states of the memory cell and further update the cells, hence enabling the network to exploit temporal information [16]. An LSTM cell as represented in Figure 2 with input feature x_t takes input data x, at time t, so that an input gate regulates the flow of the input data to the cell. A forget gate regulates when to forget contents of the internal state of the cell, and the output gate governs flow to the output. This architecture permits modeling of complicated temporal dynamics in patient health status.

2.3.2. Bidirectional LSTM architecture

The Bidirectional Long Short-Term Memory (BiLSTM) combines two parallel LSTM layers to form a forward and backward loop, as seen in Figure 3. The goal is for the network to take advantage of previous and future information through the forward and backward sequences to generate predictions. In this situation, current information has previous information as dependencies and also related to future information. [16] The forward and backward sequences respectively are illustrated by the gray and green arrows in Figure 3.

2.3.3. Bidirectional LSTM with Attention mechanism

We constructed a deep learning architecture utilizing Bidirectional Long Short-Term Memory (BiLSTM) networks enhanced by an attention mechanism, as depicted in Figure 4. The design processes 28-dimensional feature vectors via a hierarchical structure consisting of three recurrent layers succeeded by completely connected layers. The input features are initially reshaped and subsequently processed through two stacked BiLSTM layers, including 128 and 64 units per direction, respectively. Each layer is regularized using dropout (0.4) and recurrent dropout (0.2), followed by batch normalization to enhance training stability. A scaled dot-product attention method is subsequently employed to extract the most pertinent temporal features from the BiLSTM outputs, enabling the model to concentrate on essential patterns within the sequence data.

Table 1. Performance metrics and mathematical definitions.

Metric	Formula	Description
Accuracy	$\frac{(TP + TN)}{(TP + TN + FP + FN)}$	Overall correctness of the model; proportion of all correct predictions
Precision	$\frac{TP}{(TP + FP)}$	Of all positive predictions, how many were actually positive
Recall (Sensitivity)	$\frac{TP}{(TP + FN)}$	Of all actual positives, how many were correctly identified
Specificity	$\frac{TN}{(TN + FP)}$	Of all actual negatives, how many were correctly identified
F1-Score	$2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$	Harmonic mean of precision and recall; balances both metrics
MCC	$\frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	Correlation between predicted and actual classes; handles imbalanced data well
AUC-ROC	Area under ROC curve, which plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at various classification thresholds.	Measures ability to distinguish between classes across all thresholds; 1.0 = perfect, 0.5 = random

The attention-weighted representations undergo processing via a unidirectional LSTM layer of 32 units, succeeded by two dense layers containing 64 and 32 neurons, respectively, employing ReLU activation and dropout regularization. A sigmoid-activated output neuron generates the binary classification prediction. This architecture integrates the bidirectional context modeling of BiLSTM layers, the selective emphasis of attention mechanisms, and the regularization advantages of dropout and batch normalization to attain strong performance in the classification problem. The architecture contains approximately 528,000 trainable parameters.

2.3.4. Training Configuration

The model utilized the Adam optimizer with a learning rate of 0.001 and employed binary cross-entropy as the loss function, processing data in batches of 32 samples. The training was set for a maximum of 50 epochs, including various regularization techniques to mitigate overfitting and improve generalization. In addition to the dropout and batch normalization layers incorporated into the architecture, we utilized early stopping with a patience of 15 epochs to terminate training when validation performance stagnated, as well as a learning rate reduction callback with a patience of 7 epochs to adaptively modify the learning rate upon reaching performance plateaus. This thorough training setting facilitated rapid convergence while preserving model resilience and averting overfitting to the training data.

2.4. Baseline models for comparison

To assess the efficacy of our proposed BiLSTM-attention architecture, we evaluated its performance against a variety of baseline models, including both conventional machine learning and alternative deep learning methodologies. The conventional machine learning baselines comprised Logistic Regression [26] with L2 regularization for linear classification benchmarking, Random Forest [27] with 100 trees to encapsulate non-linear relationships via ensemble learning, and XGBoost [28] with 100 boosting iterations to utilize gradient boosting for improved predictive efficacy. Furthermore, we employed alternative deep learning architectures, including a conventional LSTM network [17] devoid of attention mechanisms to evaluate the impact of the attention layer, and a 1D Convolutional Neural Network [29] to assess the efficacy of convolutional feature extraction in contrast to recurrent processing. This thorough comparison allowed us to illustrate the merits of our proposed architecture across several modeling paradigms and emphasize the distinct advantages of integrating bidirectional recurrent processing with attention mechanisms.

2.5. Model training and validation

To ensure an in-depth evaluation of our proposed architecture, we established a rigorous training and validation system aimed at assessing model performance across various dimensions while mitigating overfitting and data leaking. This method integrated rigorous cross-validation techniques with meticulously chosen performance criteria to deliver a comprehensive evaluation of categorization efficacy.

2.5.1. Cross-Validation strategy

Stratified 10-fold cross-validation was utilized as the principal validation strategy to guarantee accurate performance assessment across various data subsets. To uphold the integrity of the validation process and avert data leakage, SMOTE-ENN was implemented just on the training partitions of each fold, guaranteeing that the synthetic

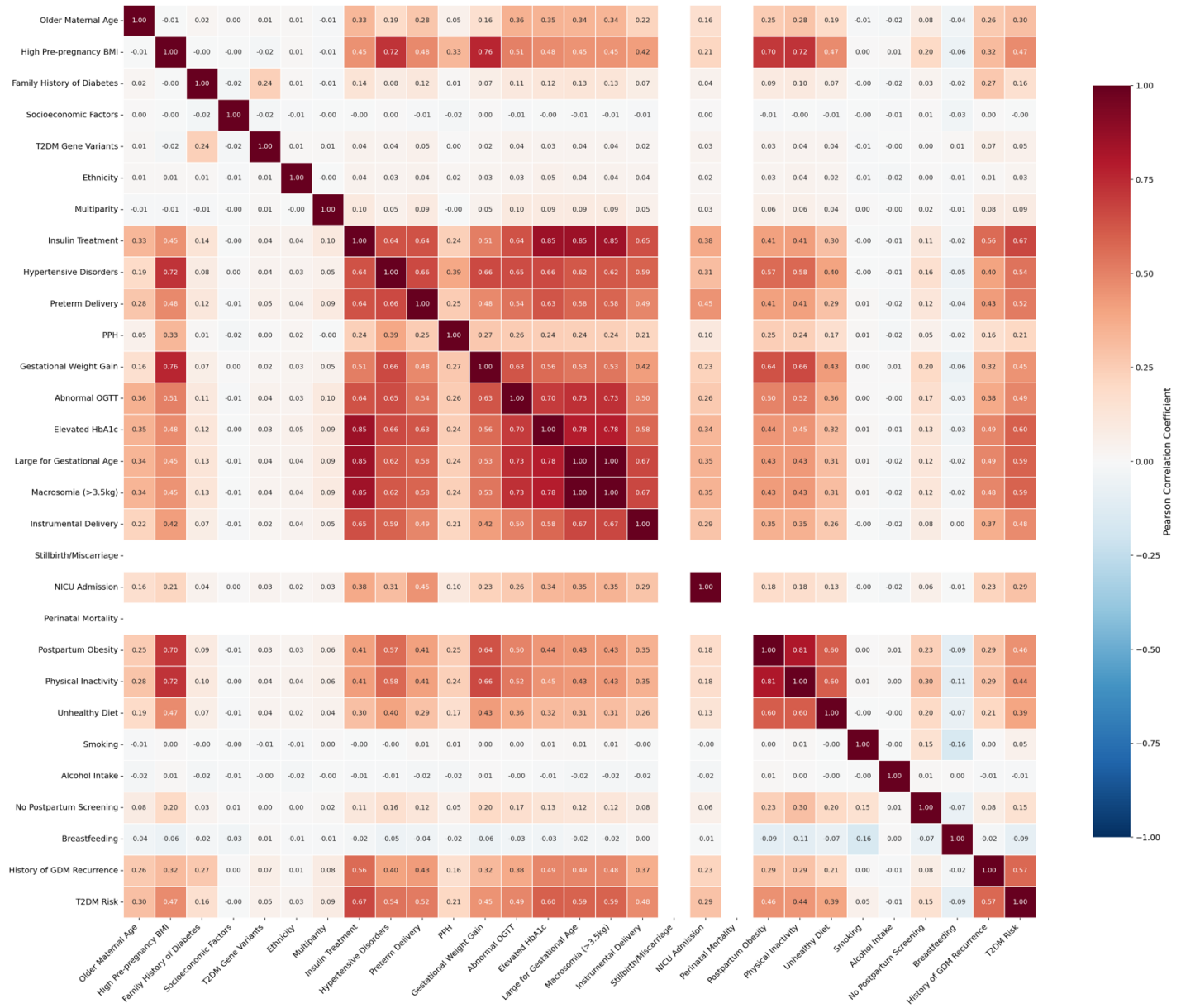

Figure 5. Correlation Heatmap.

oversampling and under-sampling methods did not compromise the validation sets. This method enabled us to tackle class imbalance while maintaining the independence of test data in each fold.

2.5.2. Performance metrics

The model's performance was assessed utilizing a hierarchical array of assessment measures to encapsulate several dimensions of classification quality. Accuracy and AUC-ROC functioned as the principal metrics, offering comprehensive assessments of proper categorization and discriminative capability across diverse decision thresholds. Secondary measures like precision, recall, F1-score, Matthew's correlation coefficient (MCC), and specificity supplemented this analysis, offering a thorough evaluation of the model's performance concerning true positives, false positives, true negatives, and false negatives. The mathematical definitions of each performance metric are provided in Table 1, where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

2.6. Statistical significance testing

We performed thorough hypothesis testing on the cross-validation results to ascertain the statistical validity of performance disparities between our suggested model and baseline techniques. Paired t-tests and Wilcoxon signed-rank tests were undertaken to evaluate the statistical significance of observed performance enhancements, with all analyses performed at a significance level of α = 0.05. Cohen's d effect sizes were computed to measure the extent of performance disparities, offering insight into both the statistical significance and practical relevance of the gains.

2.7. Explainable AI for interpretability

The study utilized the SHAP technique of explainable artificial intelligence to provide global and localized explanations for predictions and feature relationships. SHAP

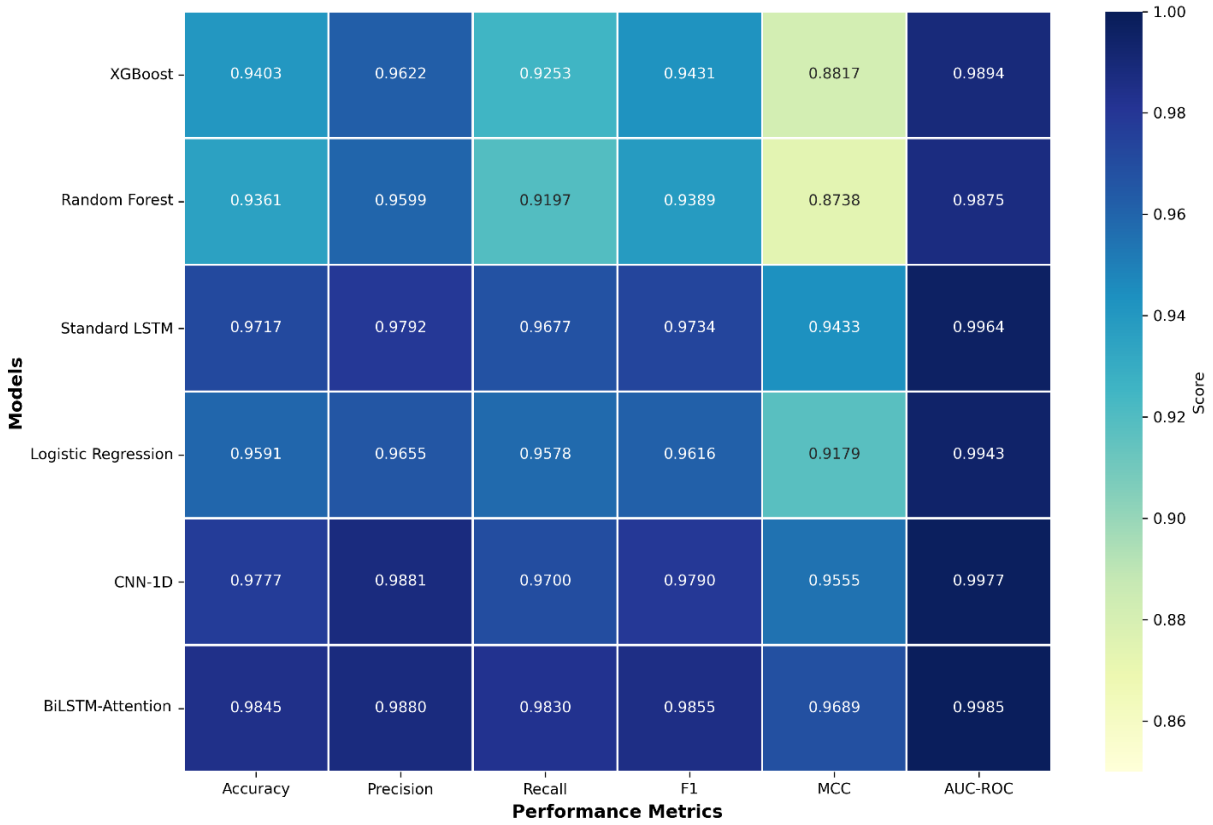

Figure 6. Model performance heatmap – mean scores.

Table 2. Comprehensive performance comparison of machine learning and deep learning models for T2DM risk prediction.

Model	Accuracy	Precision	Recall	F1	MCC	AUC-ROC
XGBoost	0.9403 ± 0.0101	0.9622 ± 0.0173	0.9253 ± 0.0201	0.9431 ± 0.0096	0.8817 ± 0.0200	0.9894 ± 0.0034
Random Forest	0.9361 ± 0.0101	0.9599 ± 0.0183	0.9197 ± 0.0277	0.9389 ± 0.0103	0.8738 ± 0.0195	0.9875 ± 0.0029
Standard LSTM	0.9717 ± 0.0068	0.9792 ± 0.0077	0.9677 ± 0.0076	0.9734 ± 0.0064	0.9433 ± 0.0138	0.9964 ± 0.0016
Logistic Regression	0.9591 ± 0.0063	0.9655 ± 0.0080	0.9578 ± 0.0092	0.9616 ± 0.0060	0.9179 ± 0.0127	0.9943 ± 0.0017
CNN-1D	0.9777 ± 0.0040	**0.9881 ± 0.0035**	0.9700 ± 0.0067	0.9790 ± 0.0039	0.9555 ± 0.0080	0.9977 ± 0.0007
BiLSTM-Attention	**0.9845 ± 0.0039**	0.9880 ± 0.0032	**0.9830 ± 0.0055**	**0.9855 ± 0.0036**	**0.9689 ± 0.0077**	**0.9985 ± 0.0006**

employs Shapley values from game theory to clarify feature significance in machine learning models. Advantages include the supply of contrasting explanations, a solid theoretical basis, and thorough clarifications evenly distributed among feature values. feature values.

2.8. Ethical Considerations

This research employed a publicly accessible synthetic dataset of entirely de-identified simulated records. Due to the lack of actual data, a formal Institutional assessment Board assessment was not mandated according to laws governing research with publicly available, non-human-subject data. All research protocols complied with ethical standards for responsible AI development in healthcare.

3. Results

3.1 Correlation Analysis

Figure 5 provides a Pearson correlation heatmap displaying the pairwise correlations among all features in the dataset. The study reveals numerous notable relationships important to T2DM risk prediction. Strong positive associations were detected across pregnancy-related problems, particularly amongst Elevated HbA1c, Abnormal OGTT findings, Large for Gestational Age, and Macrosomia ($r > 0.60$), suggesting these variables typically co-occur in high-risk pregnancies. Lifestyle factors including Physical Inactivity, Unhealthy Diet, and Postpartum Obesity revealed modest positive relationships with each other ($r = 0.40–0.55$), showing clustering of modifiable risk behaviors. The target variable, T2DM Risk, showed the highest connec-

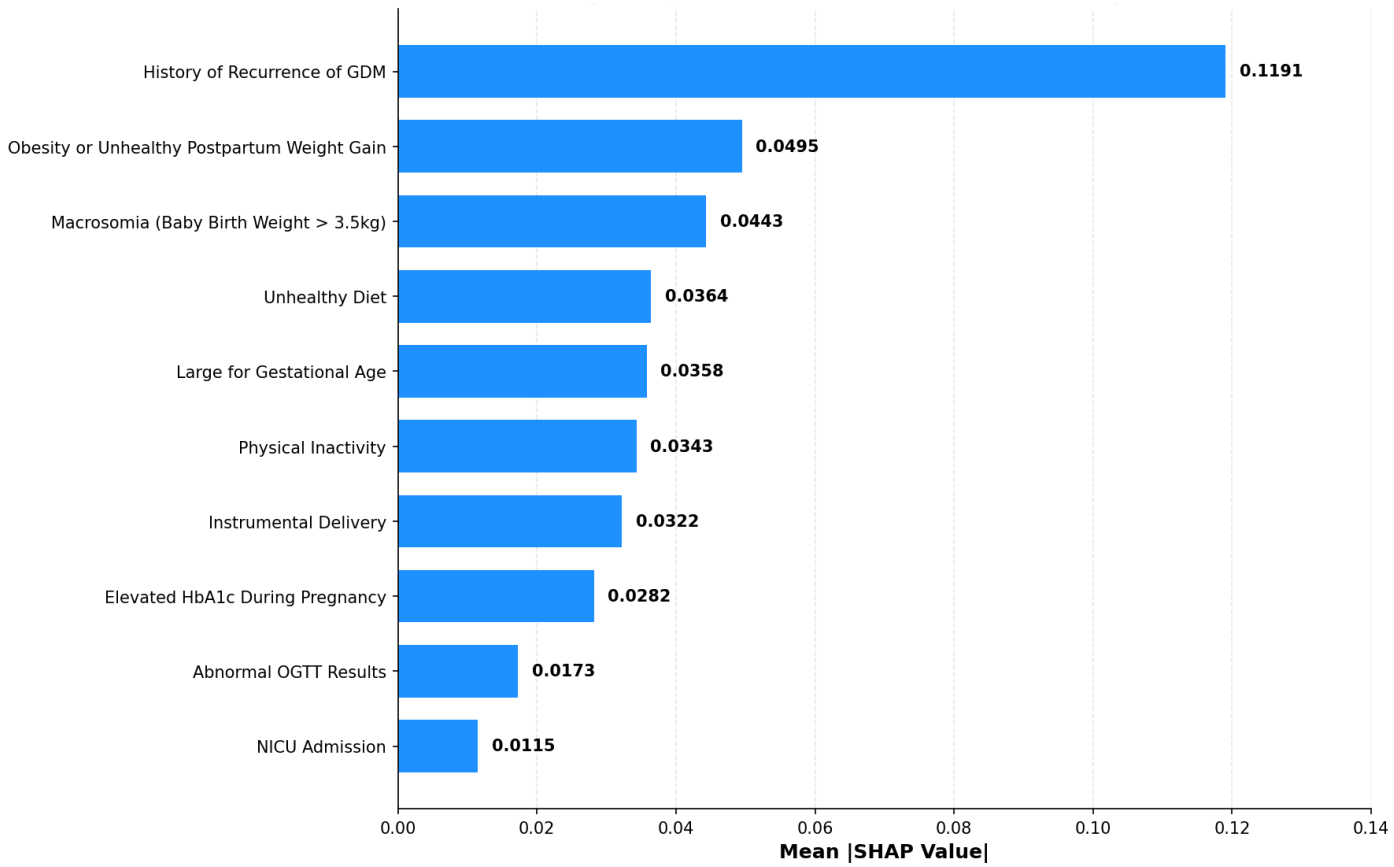

Figure 7. Top 10 features by Mean Absolute SHAP value.

Table 3. Statistical significance testing.

Comparison	Metric	t-statistic	p-value	Significance	Effect Size (d)
vs. Random Forest	ACCURACY	16.162	0.0000	p<0.001	5.804
	AUC_ROC	9.312	0.0000	p<0.001	3.656
	F1	15.814	0.0000	p<0.001	5.834
	MCC	16.312	0.0000	p<0.001	5.755
vs. XGBoost	ACCURACY	10.740	0.0000	p<0.001	3.652
	AUC_ROC	7.697	0.0000	p<0.001	2.597
	F1	10.744	0.0000	p<0.001	3.706
	MCC	10.598	0.0000	p<0.001	3.593
vs. Standard LSTM	ACCURACY	8.031	0.0000	p<0.001	2.234
	AUC_ROC	6.944	0.0001	p<0.001	1.805
	F1	8.254	0.0000	p<0.001	2.313
	MCC	7.692	0.0000	p<0.001	2.165
vs. Logistic Regression	ACCURACY	13.903	0.0000	p<0.001	4.880
	AUC_ROC	8.895	0.0000	p<0.001	3.326
	F1	13.620	0.0000	p<0.001	4.898
	MCC	13.977	0.0000	p<0.001	4.845
vs. CNN-1D	ACCURACY	1.146	0.2812	not significant	0.419
	AUC_ROC	1.340	0.2131	not significant	0.602
	F1	1.259	0.2396	not significant	0.423
	MCC	1.043	0.3243	not significant	0.398

tion with History of GDM Recurrence (r = 0.32), followed by Postpartum Obesity (r = 0.28) and Insulin Treatment during pregnancy (r = 0.26). Notably, most features displayed low to moderate intercorrelations (r < 0.50), suggesting minimal multicollinearity concerns for the predictive models. These connection patterns correspond with recognized clinical understanding that recurrent GDM and postpartum metabolic variables are significant drivers in the development from GDM to T2DM.

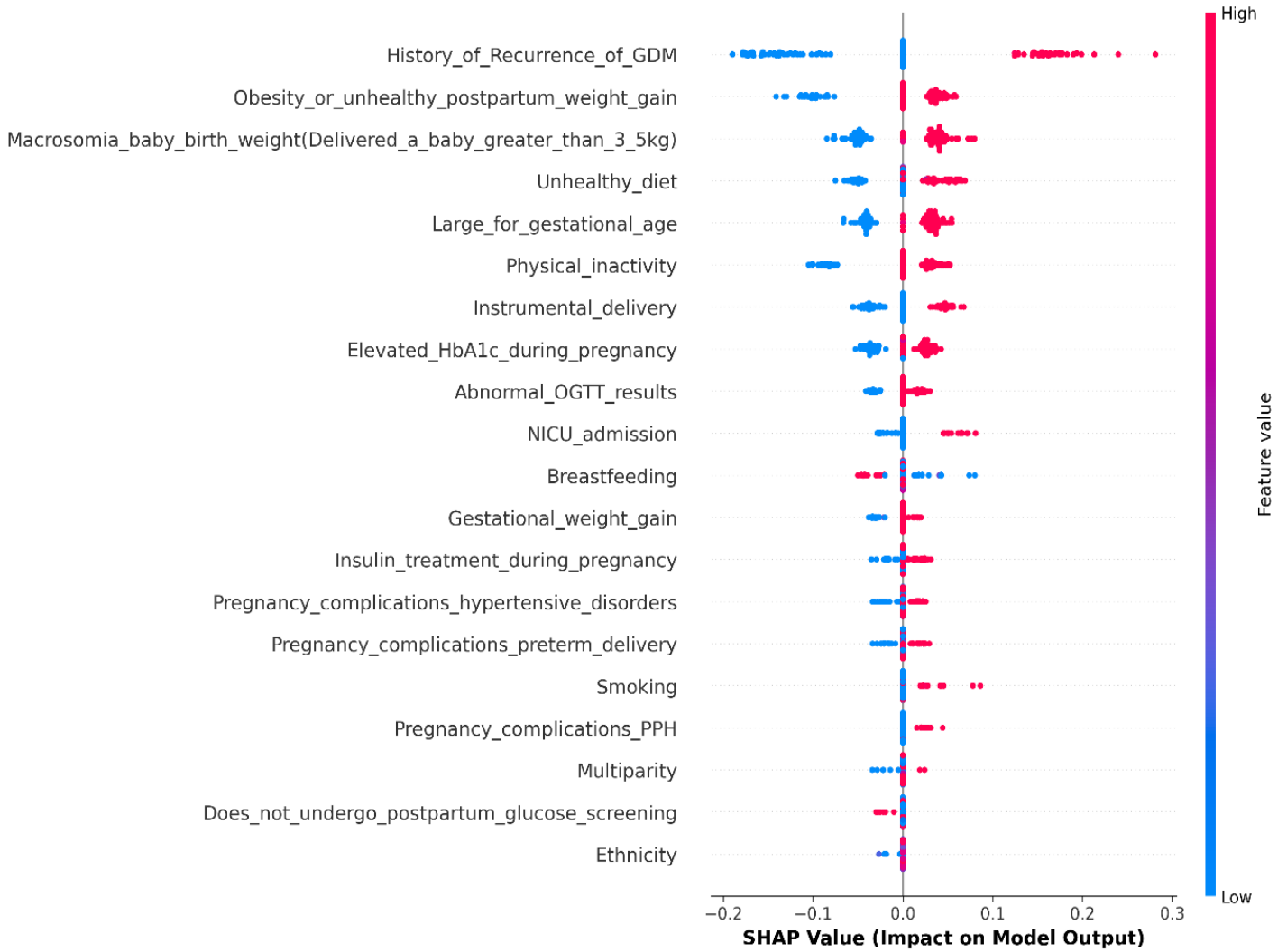

Figure 8. SHAP feature importance plot for the model's predictions.

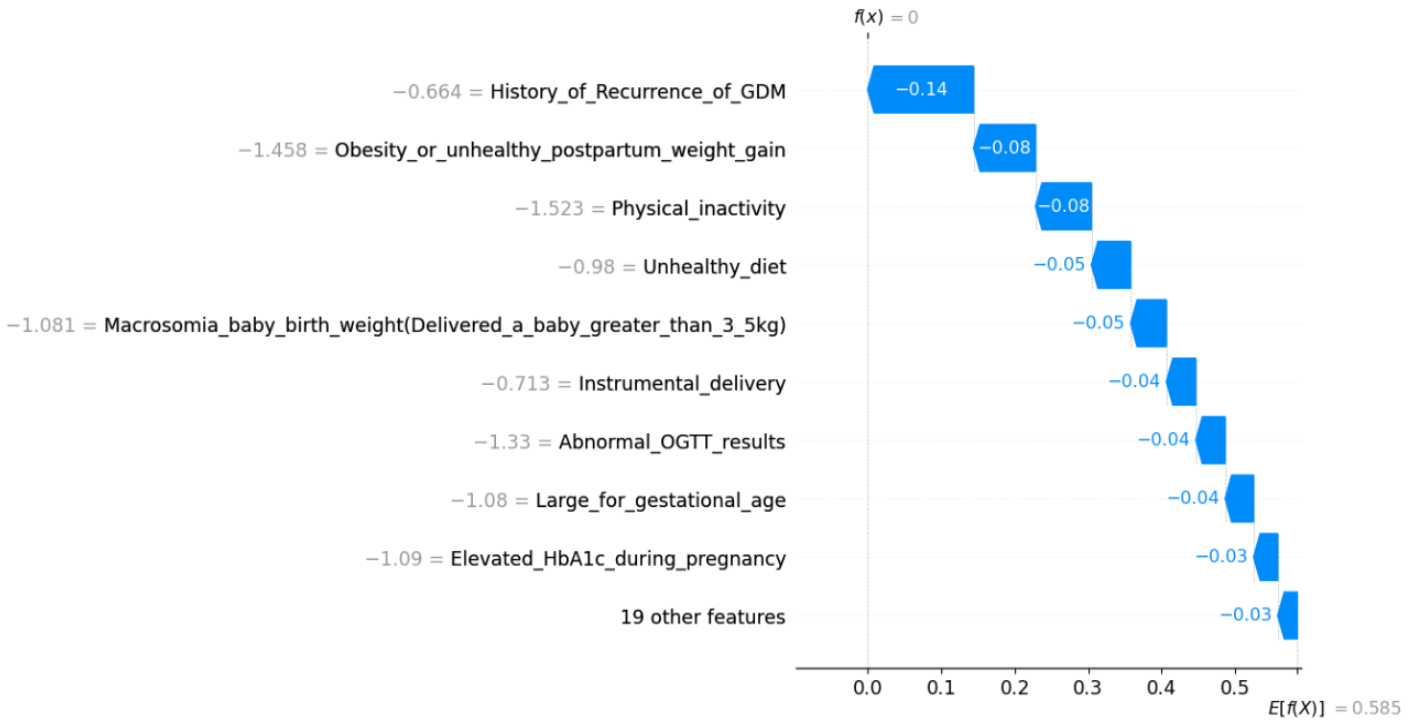

Figure 9. SHAP waterfall plots for low-risk case.

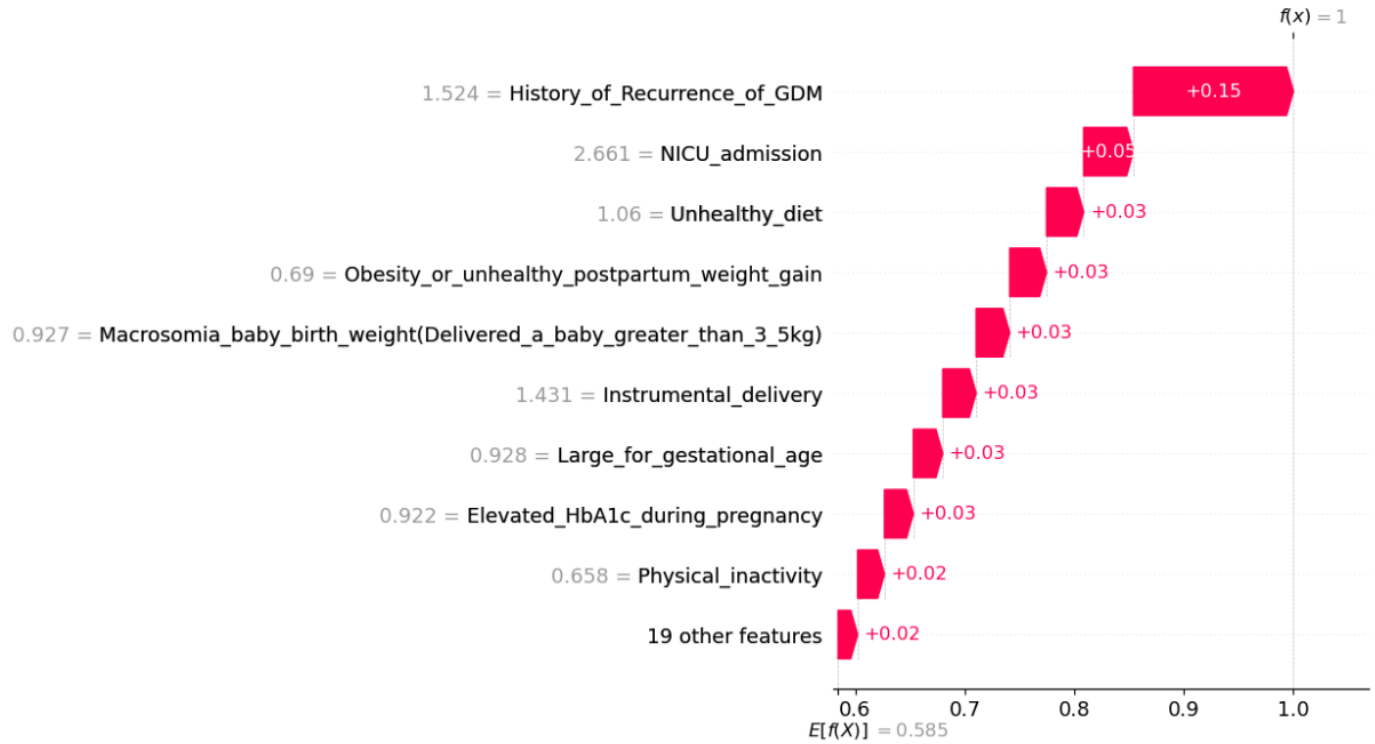

Figure 10. SHAP waterfall plots for high-risk case.

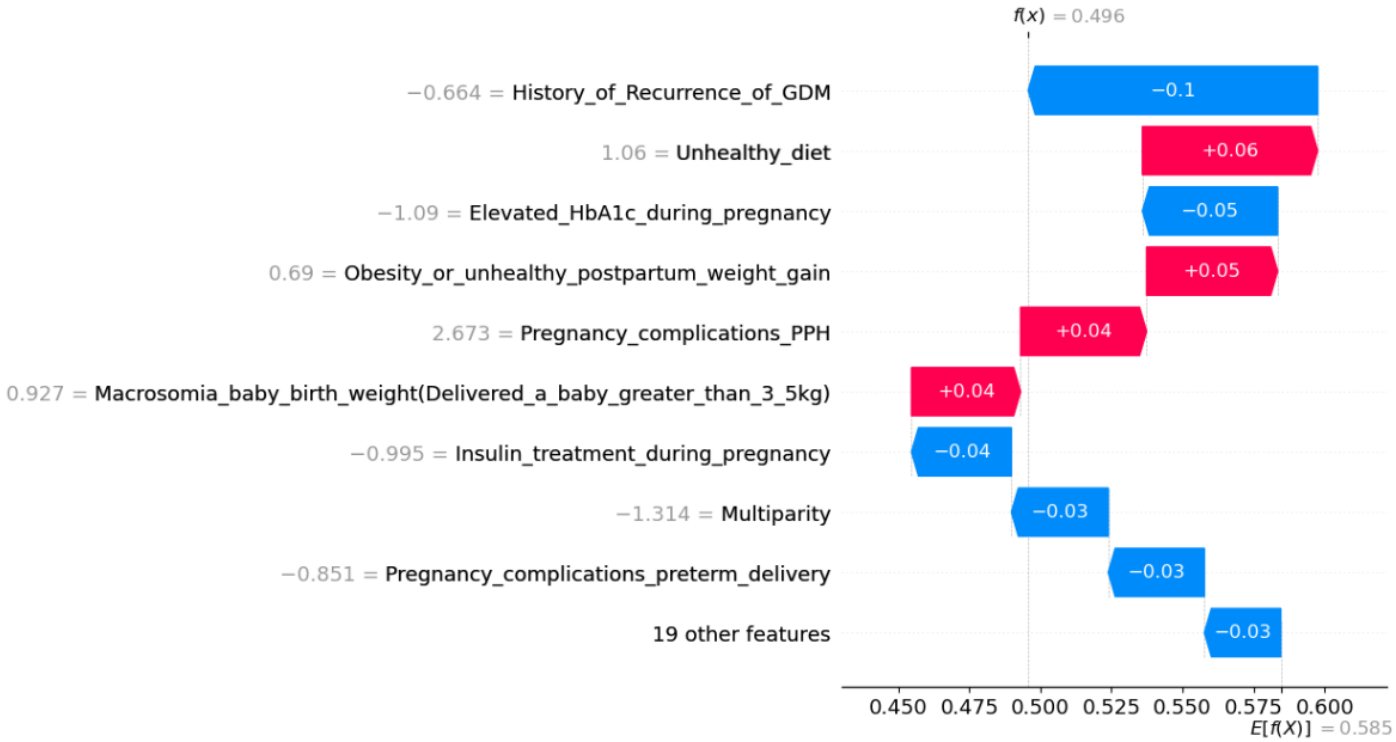

Figure 11. SHAP waterfall plots for borderline case.

3.2. BiLSTM-Attention performance

The thorough comparison demonstrates BiLSTM-Attention's outstanding performance across all evaluation parameters, notably excelling in precision (98.80%), which is essential for reducing false positive predictions in clinical environments. The model exhibits an exceptional equilibrium between precision and recall (98.30%), attaining an F1 score of 98.55% that markedly surpasses other baseline methodologies. Table 2 presents a comprehensive performance comparison of machine learning and deep learning models for T2DM risk prediction. Figure 6 depicts a heatmap visualization of the mean performance scores obtained across different evaluation metrics.

3.3. Statistical significance analysis

The paired t-test analysis demonstrates that BiLSTM-Attention significantly outperforms all competing models except CNN-1D across all key performance metrics. The

most substantial performance advantage is observed against Random Forest, with exceptionally large t-statistics ranging from 9.312 to 16.312 and effect sizes between d = 3.656 and d = 5.834, indicating massive practical significance ($p < 0.001$). BiLSTM-Attention also shows highly significant superiority over XGBoost (t-values: 7.697-10.744, d = 2.597-3.706) and Logistic Regression (t-values: 8.895-13.977, d = 3.326-4.898), with all comparisons achieving $p < 0.001$. The comparison with Standard LSTM reveals significant but somewhat smaller differences (t-values: 6.944-8.254, d = 1.805-2.313, $p \leq 0.0001$), still representing meaningful performance gains. Notably, when compared to CNN-1D, BiLSTM-Attention shows no statistically significant differences ($p > 0.05$) across all metrics, with small effect sizes ($d < 0.61$), suggesting these two models perform comparably. Statistical significance testing results are illustrated in Table 3.

However, BiLSTM-Attention achieves the highest mean AUC-ROC score of 0.9985 ± 0.0006, establishing it as the best-performing model overall. These results provide compelling statistical evidence that BiLSTM-Attention represents the superior choice for this classification task, demonstrating exceptional performance with the largest effect sizes observed against Random Forest and consistently outperforming traditional machine learning and deep learning approaches.

3.4. SHAP-Based model interpretability

SHAP employs Shapley values from game theory to provide local or global elucidations concerning feature significance for any machine learning model [30], [31]. Advantages encompass SHAP enables contrastive explanations, is underpinned by strong theoretical foundations, and offers thorough explanations that are fairly distributed among feature values. SHAP requires significant processing resources [32].

3.4.1. Global feature-importance

The Shapley value is defined as the marginal contribution of a variable's value to prediction over all potential "coalitions" or subsets of features [30]. Top 10 features by Mean Absolute SHAP value are illustrated in Figure 7.

Figure 8 illustrates the SHAP feature importance values for predicting the risk of T2DM. Each point signifies a distinct prediction, with the x-axis illustrating the SHAP value (influence on model output) and the y-axis enumerating features ordered by significance. The color gradient signifies feature values, with red denoting high values and blue indicating low values. Features are ranked from most to least influential, with History of recurrence of GDM exhibiting the greatest impact on predictions.

3.4.2. Individual prediction explanations

A SHAP waterfall plot depicts the expected SHAP values for a specific sample, including all dimensions. The model properties are oriented along the y-axis, with each unique sample's corresponding value indicated in gray. The SHAP value for each feature associated with this specific sample is displayed in the main panel, indicated by an arrow for each row or feature. A row is designated as red (blue) if the SHAP value elevates (diminishes) the prediction f(xi) relative to the anticipated or mean projection. This plot type offers a localized interpretation by concentrating on a singular sample. Figure 9-11 depicts the results of a SHAP waterfall plot following prediction.

4. Discussion

4.1. Primary findings

This study illustrates the methodological viability of utilizing a bidirectional LSTM neural network with an attention mechanism, integrated with SHAP explainability, for predicting T2DM risk in women with a history of GDM. By employing a meticulously designed synthetic dataset that mirrors epidemiological trends from existing literature, we attained remarkable discriminative efficacy (AUC-ROC: 0.9968) while ensuring transparent and interpretable predictions. This analysis yielded three principal findings. The BiLSTM-Attention architecture significantly outperformed all baseline models, demonstrating large to very high effect sizes and showing its technical superiority in the synthetic data context. Secondly, SHAP analysis identified recognized clinical risk factors and uncovered actionable modifiable factors, illustrating that high predictive performance can be attained without compromising interpretability. The model exhibited remarkable stability (AUC-ROC SD: 0.0021), indicating reliable performance within the synthetic data context and uniformity across various data subsets.

4.2. Methodological contributions

This study introduces multiple methodological advancements that enhance the domain of AI-driven diabetes prediction. This is the inaugural application of the BiLSTM-Attention architecture specifically for predicting post-GDM diabetes, hence broadening the scope of deep learning approaches beyond conventional uses. Secondly, we incorporated SHAP explainability from the outset of development instead of as an afterthought, guaranteeing that interpretability was a fundamental design feature. Third, we performed thorough comparative validation against various baseline models, offering clear benchmarking that situates our method within the current methodological framework.

4.3. Potential clinical implications (Following real-world validation)

Subject to successful validation on actual clinical data, this methodology may facilitate significant enhancements in postpartum care for women with a history of gestational diabetes mellitus (GDM). Risk-stratified screening tech-

niques may supplant uniform approaches, allowing for more regular monitoring of high-risk women while alleviating superfluous testing for those at reduced risk. Personalized intervention planning may focus on modifiable risk factors found via SHAP analysis, hence formulating individualized prevention plans instead of generic lifestyle advice. Patient-centered explainable forecasts may improve collaborative decision-making by offering transparent justifications for risk evaluations that both patients and clinicians can comprehend and utilize. Optimized resource allocation could channel limited healthcare resources to people most likely to benefit, potentially enhancing both efficiency and equity in care delivery. Nonetheless, it is crucial to highlight that all these prospective advantages remain conjectural until substantiated by thorough real-world investigations and prospective trials that demonstrate genuine clinical efficacy and patient outcomes.

4.4. Limitations

Being dependent on synthetic instead of real clinical data constitutes the principal limitation of this investigation. The synthetic dataset, although meticulously crafted to mirror established epidemiological trends, fails to encapsulate the complete intricacy, diversity, noise, and confounding factors inherent in actual clinical settings. The outstanding performance metrics (AUC-ROC: 0.9968) presumably indicate an upper limit that may not be attainable when the model is utilized on real clinical data, which is characterized by its intrinsic complexity and ambiguity. The restricted external validity indicates that although these findings illustrate methodological feasibility on controlled data, they cannot be extrapolated to clinical practice without comprehensive real-world validation across varied patient demographics and healthcare environments. The reduced clinical reality inherent in synthetic generation inevitably lowers the intricate biological and social processes contributing to diabetes development to mathematical equations that may not accurately represent genuine causative mechanisms. Moreover, the established ground truth in synthetic data establishes deterministic correlations between features and outcomes that are absent in actual clinical predictions, where outcomes are intrinsically uncertain and affected by unmeasured variables, potentially resulting in significant overestimation of discriminative capability.

Various methodological selections place supplementary limitations on interpretation. Binary feature encoding led to the loss of detailed information from continuous observations, potentially eliminating predictive signals. The elevated parameter count in relation to sample size creates apprehensions regarding possible overfitting, notwithstanding the implementation of cross-validation techniques. The application of SMOTE+ENN to rectify class imbalance may have generated synthetic oversampling artifacts that distort performance metrics. The absence of genuine longitudinal dynamics prevents the model from accurately capturing the temporal patterns of risk factor evolution that could be clinically significant.

Significant obstacles to clinical translation persist unresolved. The absence of real-world validation indicates that the model has not been evaluated on actual patients in clinical environments. The obstacles to implementation, including as issues in EHR integration, workflow disturbance, computing demands, and regulatory approval processes, remain unexamined. The generalizability across diverse demographics, healthcare systems, and geographic regions is currently uncertain. The clinical value of this approach—specifically, its impact on patient outcomes—remains unknown and necessitates prospective randomized trials for validation.

This study should be regarded as a methodological proof-of-concept illustrating computational feasibility rather than a therapeutically viable instrument. Clinical adoption necessitates stringent real-world validation, prospective trials evidencing enhanced results, and regulatory endorsement prior to any patient care applications.

4.5. Future research directions

The key subsequent step is the authentication of authentic clinical data from various healthcare systems. This validation must evaluate actual performance metrics on authentic patient records, calibration across various risk strata and subpopulations, temporal stability as clinical practices and populations change, and direct comparison with clinical judgment and existing risk assessment instruments. In the absence of this validation, all further developments are merely theoretical endeavors lacking practical significance. Conducting multi-site validation across various healthcare systems, geographic areas, and patient demographics is crucial for establishing generalizability and identifying potential sources of performance variation or algorithmic bias. Furthermore, augmenting the existing system to include temporal modeling for time-to-event prediction will allow doctors to derive accurate timeline estimates for T2DM onset, thereby enabling more strategically timed intervention approaches.

4.6. Comparative performance analysis

The performance of our model on synthetic data (AUC-ROC: 0.9968) much surpasses that of previously reported models in real-world post-GDM populations (AUC-ROC: 0.70-0.92). This comparison should be approached with some caution, as performance on synthetic data likely indicates an upper limit that may not be attainable with actual clinical records due to intrinsic noise, missing data, and unmeasured confounders. The synthetic data environment mitigates numerous real-world obstacles that commonly constrain model efficacy, such as measurement inaccuracies, insufficient documentation, inconsistent data quality across healthcare systems, and

Table 4. Comprehensive comparison of previous studies for T2DM risk prediction with GDM history.

Study	Dataset Size	Data Type	Scope	Key Features	Best Algorithm	Performance Metrics
[33]	N/A	Synthetic	Test AIRS algorithm for predicting GDM to T2D transition on im-balanced datasets	• Imbalanced dataset • Clinical and demographic features	AIRS	• Classification recall: 62.8% • Better than LogReg & SVM
[34]	104	Real-world + Omics	Assess if lipidomic profiling enhances T2DM risk prediction in women post-GDM (median 8.5 years follow-up)	• Age, BMI • Pregnancy fasting glucose • Postnatal fasting glucose • Triacylglycerol, total cholesterol • CE 20:4, PE(P-36:2), PS 38:4	Clinical prediction model	• Net reclassification index: +22.3% • Three lipid species significantly associated with T2D progression
[35]	257	Real-world	Create simple, easy-to-calculate risk score for long-term diabetes risk after GDM (20-year prospective follow-up)	• BMI in early pregnancy • Insulin treatment during pregnancy • Family history of diabetes • Lactation duration	Lasso Cox regression	• R^2: 0.23-0.33 • C-index: 0.75
[5]	1,035	Real-world + Omics	Develop metabolomics signature to predict T2DM from single fasting blood sample at 6-9 weeks postpartum	• 21 metabolites identified via metabolomics • Baseline fasting plasma	Decision Tree	• Training set accuracy: 83.0% • Testing set accuracy: 76.9% • AUC-ROC: 0.77
[4]	140 (review) / 1,035 (additional)	Real-world + Omics	Discover novel predictive biomarkers and early-stage pathophysiology for GDM to T2DM transition	• 7 lipid metabolites (review) • 21 lipolytic metabolites (additional) • Sphingolipid metabolism markers	Decision Tree	• AUC-ROC: 0.92 • Accuracy: 91% • Sensitivity: 87% • Specificity: 93%
[36]	1,263	Real-world	Develop nomogram for incident risk of postpartum T2DM using non-invasive clinical characteristics	• Family history of diabetes • Pregnancy-induced hypertension • Pre-pregnancy BMI • 2-hour glucose at 26-30 weeks	Multivariate Cox proportional hazards model	• AUROC: 82.8% (95% CI: 78.1%-87.5%) • 2-year AUROC: 85.9% • 3-year AUROC: 83.2%
[37]	1,035	Real-world + Omics	Identify metabolic signature in early postpartum period (6-9 weeks) predic-	• Fasting plasma metabolites • Amino acids	Not specified	• Median AUC: 0.883 • 95% CI: 0.820-0.945

			ting T2DM transition	• Diacyl-glycero-phospholipids • Sphingolipids • Acyl-alkyl-glycero-phospholipids		• $p < 0.001$
[6]	103 (review) / 754 (additional)	Real-world + Omics	Evaluate if postpartum circulating miRNAs enhance T2DM prediction beyond traditional clinical factors	• 754 plasma circulating miRNAs • miR-369-3p (key biomarker) • Age, BMI, fasting glucose, lipids	Penalized LogReg + bootstrapping	• AUC: 0.92 (with miRNAs) • AUC: 0.83 (clinical only) • Sensitivity: 91% • Specificity: 89%
[38]	317	Real-world	Develop clinical diabetes risk prediction model for prediabetic women with prior GDM from DPP study	• 11 baseline clinical variables • Final model: 4 variables (fasting glucose, HbA1c, BMI, treatment arm)	Cox proportional hazards regression	• C-index: 0.68 (bias-corrected)
[8]	692	Real-world	Quantify T2DM and dysglycemia risk using pre-pregnancy and pregnancy factors	• GDM status • Pre-pregnancy BMI • PDWR	Poisson Regression	• RR (T2DM): 12.07 (95% CI: 4.55-32.02) • RR (Dysglycemia): 3.02 (95% CI: 1.14-7.98)
[9]	561	Real-world	Predict T2DM risk using midpregnancy clinical features	• Midpregnancy BMI • GDM diagnosis	CatBoost	• AUC-ROC: 0.86 • 95% CI: 0.72-0.99
[39]	6,092	Real-world	Use ML to predict T2DM based on glucose metabolism patterns during pregnancy	• Age, parity, gravidity • GCT and OGTT results • Gestational age at delivery • Birthweight	XGBoost	• AUC: 0.85 • Accuracy: 91% • Sensitivity: 74% • Specificity: 74%
[40]	78	Real-world	Identify key factors determining T2DM development in women with GDM history using ML techniques	• Age, BMI • Fasting glucose • Insulin secretion/ action indicators • (34 features → 6 after selection)	L2-penalized LogReg	• AUC: 0.884 (LogReg) • AUC: 0.831 (Naïve Bayes) • AUC: 0.795 (Decision tree) • Accuracy: 84.0% • F1-Score: 0.828-0.836
[41]	607	Real-world	Predict non-attendance at postpartum glucose screening and subsequent T2DM risk	• Antenatal factors: age, parity • BMI, smoking status • Fasting glucose at OGTT	ML model (not specified)	• AUC: 0.72 • Sensitivity: 70% • Specificity: 66%

[42]	1,299	Real-world	Develop antenatal and postnatal risk prediction models for T2DM in women with GDM enrolled in LIVING study	• Glucose test results • Medical history • Biometric indicators	Not explicitly stated	• Antenatal AUC: 0.76 (95% CI: 0.72-0.80) • Postnatal AUC: 0.85 (95% CI: 0.81-0.88) • Accuracy: 70.82% (antenatal), 76.10% (postnatal)
[7]	6,000	Synthetic	Evaluate Fundamental and ensemble ML methods for T2DM prediction using comprehensive risk factors	• 28 clinical risk factors including Maternal characteristics, Genetic risk factors and Lifestyle factors	AdaBoost	• F1-Score: 80.4%
This study	6,000	Synthetic	BiLSTM-Attention for high-accuracy T2DM prediction with comprehensive features	• 28 clinical risk factors including Maternal characteristics, Genetic risk factors and Lifestyle factors • Used SMOTE-ENN for balancing	BiLSTM-Attention	• AUC-ROC: 0.9986 • Accuracy: 98.45% • Precision: 98.80% • Recall: 98.30% • F1-Score: 98.55% • MCC: 96.89%

the intricate interactions of unmeasured social and biological variables. Consequently, although these findings illustrate the theoretical capabilities of the BiLSTM-Attention architecture, they should not be construed as proof of its superiority over current models until equivalent validation on actual clinical data is conducted. Table 4 presents a comparison of prior studies regarding T2DM risk prediction in individuals with a history of GDM.

Table 4 presents a comprehensive comparison of previous studies on T2DM risk prediction in post-GDM populations, enabling evaluation of our contribution within the current literature. The comparison highlights several interesting observations: First, addressing dataset characteristics, previous research generally utilized real-world clinical data with sample sizes ranging from 78 to 6,092 participants, while our study employed synthetic data (n=6,000) specially tailored to replicate epidemiological patterns. Second, in terms of methodological approaches, prior research primarily relied on traditional machine learning methods (Logistic Regression, XGBoost, Random Forest) and statistical approaches (Cox regression, Poisson regression), whereas our BiLSTM-Attention architecture represents the first application of attention-augmented deep learning to this clinical problem. Third, addressing performance measures, our model achieved much higher AUC-ROC (0.9986) compared to real-world research showing AUC values between 0.68 and 0.92. However, this performance gap should be evaluated cautiously, as synthetic data inherently lacks the noise, missing values, and unmeasured confounders prevalent in clinical datasets. Fourth, addressing feature comprehensiveness, our model includes 28 clinical risk factors covering pre-pregnancy, pregnancy, and postpartum stages, providing one of the most comprehensive feature sets in the literature. Finally, the incorporation of SHAP explainability separates our approach by offering both global feature priority rankings and individual prediction explanations, addressing the interpretability restrictions of earlier "black box" models. These comparisons demonstrate that while our BiLSTM-Attention model achieves exceptional performance on synthetic data, the true contribution lies in establishing a methodological framework that combines high-performance deep learning with explainable AI, which can be validated and refined using real-world clinical data in future studies.

5. Conclusion

This study illustrates the methodological viability of utilizing BiLSTM-Attention alongside SHAP explainability for T2DM risk prediction in women with a history of GDM. By employing meticulously crafted synthetic data, we attained outstanding performance while preserving in-

terpretability and transparency. This work serves as a proof-of-concept for building a computational framework, rather than a therapeutically viable tool for patient treatment. The integration of high-performance deep learning with explainable AI signifies a transformative change towards transparent prediction systems capable of mitigating enduring issues associated with black-box algorithms in healthcare. This study establishes a methodological framework for forthcoming real-world validation research crucial for clinical translation. The freely accessible synthetic dataset allow other researchers to replicate, evaluate, and expand upon this methodological contribution, thereby expediting innovation in this significant domain. Through rigorous validation of clinical data, prospective trials showcasing enhanced outcomes, meticulous consideration of equity and fairness, and deliberate implementation that aligns with clinical workflows, AI-driven risk stratification has the potential to revolutionize postpartum care for women with a history of gestational diabetes mellitus by facilitating earlier detection and tailored prevention strategies. This study illustrates a possible approach to achieving that objective, while recognizing the substantial validation efforts need to convert scientific advancements into quantifiable therapeutic benefits and enhanced patient outcomes.

6. Declarations

6.1. Author Contributions

Amirthanathan Prashanthan*:* Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft; **Jenifar Prashanthan:** Investigation, Formal analysis, Writing - Review & Editing, Resources, Project administration.

6.2. Institutional Review Board Statement

Not applicable.

6.3. Informed Consent Statement

Not applicable.

6.4. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.5. Acknowledgment

Not applicable.

6.6. Conflicts of Interest

The authors declare no conflicts of interest.

7. References

[1] American Diabetes Association (ADA), "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018," *Diabetes Care*, vol. 41, no. Supplement_1, pp. S13–S27, Jan. 2018, https://doi.org/10.2337/dc18-S002.

[2] International Diabetes Federation, "IDF DIABETES ATLAS: 463 million PEOPLE LIVING WITH DIABETES," 2019.

[3] E. Vounzoulaki, K. Khunti, S. C. Abner, B. K. Tan, M. J. Davies, and C. L. Gillies, "Progression to type 2 diabetes in women with a known history of gestational diabetes: Systematic review and meta-analysis," *The BMJ*, vol. 369, May 2020, https://doi.org/10.1136/bmj.m1361.

[4] S. R. Khan, H. Mohan, Y. Liu, B. Batchuluun, H. Gohil, D. A. Rijjal, Y. Manialawy, B. J. Cox, E. P. Gunderson, M. B. Wheeler, "The discovery of novel predictive biomarkers and early-stage pathophysiology for the transition from gestational diabetes to type 2 diabetes," *Diabetologia*, vol. 62, no. 4, pp. 687–703, Apr. 2019, https://doi.org/10.1007/s00125-018-4800-2.

[5] A. Allalou, A. Nalla, K. J. Prentice, Y. Liu, M. Zhang, F. F. Dai, X. Ning, L. R. Osborne, B. J. Cox, E. P. Gunderson, M. B. Wheeler, "A Predictive Metabolic Signature for the Transition From Gestational Diabetes Mellitus to Type 2 Diabetes," *Diabetes*, vol. 65, no. 9, pp. 2529–2539, Sep. 2016, https://doi.org/10.2337/db15-1720.

[6] M. V. Joglekar, W. K. M. Wong, F. K. Ema, H. M. Georgiou, A. Shub, A. A. Hardikar, M. Lappas, "Postpartum circulating microRNA enhances prediction of future type 2 diabetes in women with previous gestational diabetes," *Diabetologia*, vol. 64, no. 7, pp. 1516–1526, Jul. 2021, https://doi.org/10.1007/s00125-021-05429-z.

[7] J. Prashanthan and A. Prashanthan, "Predicting the future risk of developing type 2 diabetes in women with a history of gestational diabetes mellitus using machine learning and explainable artificial intelligence," *Prim. Care Diabetes*, Sep. 2025, https://doi.org/10.1016/j.pcd.2025.09.006.

[8] L.-W. Chen, S. E. Soh, M.-T. Tint, S. L. Loy, F. Yap, K. H. Tan, Y. S. Lee, L. P.-C. Shek, K. M. Godfrey, P. D.. Gluckman, J. G. Eriksson, Y.-S. Chong, S.-Y. Chan, "Combined analysis of gestational diabetes and maternal weight status from pre-pregnancy through post-delivery in future development of type 2 diabetes," *Sci. Rep.*, vol. 11, no. 1, p. 5021, Mar. 2021, https://doi.org/10.1038/s41598-021-82789-x.

[9] M. Kumar *et al.*, "Machine Learning–Derived Prenatal Predictive Risk Model to Guide Intervention and Prevent the Progression of Gestational Diabetes Mellitus to Type 2 Diabetes: Prediction Model Development Study," *JMIR Diabetes*, vol. 7, no. 3, p. e32366, Jul. 2022, https://doi.org/10.2196/32366.

[10] B. S. Fiskå, A. S. D. Pay, A. C. Staff, and M. Sugulle, "Gestational diabetes mellitus, follow-up of future maternal risk of cardiovascular disease and the use of eHealth technologies—a scoping review," *Syst. Rev.*, vol. 12, no. 1, p. 178, Sep. 2023, https://doi.org/10.1186/s13643-023-02343-w.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, https://doi.org/10.1038/nature14539.

[12] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016, https://doi.org/10.1001/jama.2016.17216.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, https://doi.org/10.1038/nature21056.

[14] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks.," *JMLR Workshop Conf. Proc.*, vol. 56, pp. 301–318, Aug. 2016. https://proceedings.mlr.press/v56/Choi16.pdf.

[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[16] I. K. Ihianle, A. O. Nwajana, S. H. Ebenuwa, R. I. Otuka, K. Owa, and M. O. Orisatoki, "A Deep Learning Approach for Human Activities Recognition From Multimodal Sensing Devices," *IEEE Access*, vol. 8, pp. 179028–179038, 2020, https://doi.org/10.1109/ACCESS.2020.3027979.

[17] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to Diagnose with LSTM Recurrent Neural Networks," Mar. 2017. https://doi.org/10.48550/arXiv.1511.03677.

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, https://doi.org/10.1109/78.650093.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," May 2016. https://doi.org/10.48550/arXiv.1409.0473.

[20] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York, NY, USA: ACM, Mar. 2019, pp. 275–285. https://doi.org/10.1145/3301275.3302310.

[21] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 24, Dec. 2018, https://doi.org/10.1186/s12874-018-0482-1.

[22] S. M. Lundberg *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, Oct. 2018, https://doi.org/10.1038/s41551-018-0304-0.

[23] S. M. Lauritsen *et al.*, "Explainable artificial intelligence model to predict acute critical illness from electronic health records.," *Nat. Commun.*, vol. 11, no. 1, p. 3852, Jul. 2020, https://doi.org/10.1038/s41467-020-17431-x.

[24] J. Prashanthan and A. Prashanthan, "Data for T2DM Risk prediction after GDM." Accessed: Sep. 29, 2025. [Online]. Available: https://www.kaggle.com/datasets/prashanthana/gdm-risk-data-for-t2dm-prediction.

[25] M. Lamari, N. Azizi, N. E. Hammami, A. Boukhamla, S. Cheriguene, N. Dendani, N. E. Benzebouchi, "SMOTE–ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification," in *Advances on Smart and Soft Computing*, pp. 37–49, 2021. https://doi.org/10.1007/978-981-15-6048-4_4.

[26] D. Hosmer, R. Sturdivant, and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons., 2013. https://books.google.co.id/books?id=bRoxQBIZRd4C.

[27] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, https://doi.org/10.1023/A:1010933404324.

[28] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[29] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, https://doi.org/10.1109/TNNLS.2021.3084827.

[30] S. Lundberg, "Commentary," *Ann. Emerg. Med.*, vol. 70, no. 1, pp. 30–31, Jul. 2017, https://doi.org/10.1016/j.annemergmed.2017.05.019.

[31] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges," 2020, pp. 417–431. https://doi.org/10.1007/978-3-030-65965-3_28.

[32] D. Tchuente, J. Lonlac, and B. Kamsu-Foguem, "A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications," *Comput. Ind.*, vol. 155, p. 104044, Feb. 2024, https://doi.org/10.1016/j.compind.2023.104044.

[33] H.-C. Lin, C.-T. Su, and P.-C. Wang, "An Application of Artificial Immune Recognition System for Prediction of Diabetes Following Gestational Diabetes," *J. Med. Syst.*, vol. 35, no. 3, pp. 283–289, Jun. 2011, https://doi.org/10.1007/s10916-009-9364-8.

[34] M. Lappas, P. A. Mundra, G. Wong, K. Huynh, D. Jinks, H. M. Georgiou, M. Permezel, P. J. Meikle, "The prediction of type 2 diabetes in women with previous gestational diabetes mellitus using lipidomics," *Diabetologia*, vol. 58, no. 7, pp. 1436–1442, Jul. 2015, https://doi.org/10.1007/s00125-015-3587-7.

[35] M. Köhler, A. G. Ziegler, and A. Beyerlein, "Development of a simple tool to predict the risk of postpartum diabetes in women with gestational diabetes mellitus," *Acta Diabetol.*, vol. 53, no. 3, pp. 433–437, Jun. 2016, https://doi.org/10.1007/s00592-015-0814-0.

[36] W. Li, J. Leng, H. Liu, S. Zhang, L. Wang, G. Hu, J. Mi, "Nomograms for incident risk of post-partum type 2 diabetes in Chinese women with prior gestational diabetes mellitus," *Clin. Endocrinol. (Oxf).*, vol. 90, no. 3, pp. 417–424, Mar. 2019, https://doi.org/10.1111/cen.13863.

[37] M. Lai, Y. Liu, G. V. Ronnett, A. Wu, B. J. Cox, F. F. Dai, H. L. Röst , E. P. Gunderson, M. B. Wheeler, "Amino acid and lipid metabolism in post-gestational diabetes and progression to type 2 diabetes: A metabolic profiling study," *PLoS Med.*, vol. 17, no. 5, p. e1003112, May 2020, https://doi.org/10.1371/journal.pmed.1003112.

[38] B. Man, A. Schwartz, O. Pugach, Y. Xia, and B. Gerber, "A clinical diabetes risk prediction model for prediabetic women with prior gestational diabetes," *PLoS One*, vol. 16, no. 6 June, Jun. 2021, https://doi.org/10.1371/journal.pone.0252501.

[39] O. Houri, Y. Gil, R. Chen, A. Wiznitzer, A. Hochberg, E. Hadar, A. Berezowsky, "Prediction of Type 2 Diabetes Mellitus According to Glucose Metabolism Patterns in Pregnancy Using a Novel Machine Learning Algorithm," *J. Med. Biol. Eng.*, vol. 42, no. 1, pp. 138–144, Feb. 2022, https://doi.org/10.1007/s40846-022-00685-9.

[40] L Ilari, A. Piersanti, C. Göbl, L. Burattini, A. Kautzky-Willer, A. Tura, M. Morettini, "Unraveling the Factors Determining Development of Type 2 Diabetes in Women With a History of Gestational Diabetes Mellitus Through Machine-Learning Techniques," *Front. Physiol.*, vol. 13, Feb. 2022, https://doi.org/10.3389/fphys.2022.789219.

[41] N. Periyathambi, D. Parkhi, Y. Ghebremichael-Weldeselassie, V. Patel, N. Sukumar, R. Siddharthan, L. Narlikar, P. Saravanan, "Machine learning prediction of non-attendance to postpartum glucose screening and subsequent risk of type 2 diabetes following gestational diabetes," *PLoS One*, vol. 17, no. 3, p. e0264648, Mar. 2022, https://doi.org/10.1371/journal.pone.0264648.

[42] Y. Belsti *et al.*, "Development of a risk prediction model for postpartum onset of type 2 diabetes mellitus, following gestational diabetes; the lifestyle InterVention in gestational diabetes (LIVING) study," *Clinical Nutrition*, vol. 43, no. 8, pp. 1728–1735, Aug. 2024, https://doi.org/10.1016/j.clnu.2024.06.006.