Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31360

Interaction-Enhanced and Explainable Machine Learning for Diabetes Risk Prediction

Zhiyuan Chen

School of Software Engineering, Northeastern University, Shenyang, China
20237064@stu.neu.edu.cn

Abstract. Diabetes mellitus is a common chronic metabolic disease for which early diagnosis is crucial for prevention and treatment. As the amount of structured clinical information grows, machine learning has emerged as a valuable instrument to predict diabetes risk; nevertheless, several studies demonstrate the application of machine learning models to diabetes risk prediction based on original clinical features, while highlighting a general lack of systematic inspection of feature combinations and model interpretability. Various machine learning models have been built and tested in this research to predict diabetes risk using structured clinical data. Clinically motivated interaction terms were built to capture nonlinear physiological relationships, and a two-criterion selection approach using tree-based split gain and SHAP importance was used to identify meaningful interactions. An interaction-enhanced XGBoost model was then trained and compared with baseline and complete-interaction models using conventional classification metrics. The results of the experiment indicate that the noise created by indiscriminate inclusion of interaction features can lead to degraded generalization performance and that selectively retained interactions can increase sensitivity without compromising the discriminative performance. Glucose, BMI, and age were also identified as dominant predictors and verified in diabetes prediction through feature ablation analysis. Furthermore, the SHAP interpretability analysis presented clear and clinically coherent model behavior explanations. Overall, the developed framework implies that there is a sensible trade-off between predictive efficiency and interpretability, which highlights the importance of focused feature interaction modeling of reliable predictive and explainable diabetes risk assessments.

Keywords: Diabetes prediction, Machine learning, Feature interaction, XGBoost, Model interpretability

1. Introduction

Diabetes mellitus is a common chronic metabolic disease with a rapidly increasing global prevalence. Chronic hyperglycemia can lead to severe complications, including cardiovascular disease, damage to the kidneys, ocular blindness, all of which impose a major burden on patients and healthcare systems. Therefore, individuals who are at risk must be detected early to make timely intervention possible as the disease is likely to be silent and asymptomatic. As more and more of the structured health data available in clinical examinations and electronic records undergoes analysis,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ31360

machine learning has become an essential part of the process of making the prediction of diabetes risks more accurate and efficient.

It is against this background that machine learning has emerged as a promising approach for diabetes risk assessment. Several machine learning methods have been explored in predicting diabetes. These methods have proven that data-driven models are capable of learning nonlinear correlations between physiological predictors. Rastogi and Bansal compared the classical algorithms including the Random Forest, Support Vector Machine, and Logistic Regression and found that tree-based models are robust when working with structured clinical data [1]. Khanam and Foo made a comparative study and found that traditional machine learning methods tend to be more successful than shallow neural networks when samples are small [2]. Review studies have also summarized the growing role of machine learning in diabetes prediction and highlighted the enduring limitations of imbalance in data, model generalizability, and limited interpretability of predictive models [3]. Recently, explainable artificial intelligence techniques have also come to be used by researchers to enhance model transparency, with SHAP being a popular tool in the interpretation of feature contributions in clinical prediction tasks [4,5].

Despite these developments, however, there are still a variety of research gaps. The majority of existing studies that have been implemented are based on the initial clinical variables without systematic analysis of the role of the derived interaction features that may indicate some important nonlinear physiological links. In addition, although ensemble-based models like XGBoost and LightGBM have shown high predictive accuracy, most current research pays insufficient attention to the rigorous interpretability of the model.

To fill these gaps, this study develops and analyzes several machine learning models to identify whether a patient is at risk of diabetes using a designed collection of structured clinical data. Besides popular default baselines, a set of clinically-inspired interaction features is constructed and refined using a dual-criterion selection procedure founded on split gain and SHAP significance. An XGBoost classifier has been trained on the refined version of features and its performance was compared in a systematic way with alternative model settings. Finally, SHAP-based interpretability analyses are conducted to identify the most essential variables driving model decisions and provide clinically significant facts around diabetes risk patterns. In this structure of joint modeling and explainability, the research creates a predictive system that is precise and interpretable to help the risk diagnosis of diabetes more secure and responsible by application of machine learning.

2. Methods

2.1. Data processing

The study used the Pima Indians Diabetes data consisting of eight clinical variables that are usually related to the metabolic and glycemic status. Some of the body measurements such as glucose level, blood pressure, skinfold thickness, insulin and BMI have zero values, which were not possible under normal body conditions. These values were considered as missing and filled with medians derived from the training subset to prevent information leakage.

The dataset was partitioned into training and testing sets using stratified sampling to preserve the proportion of diabetic and non-diabetic cases. Afterward all the numerical predictors were standardized with the help of calculations made on the training set, followed by the same operation on the testing data. This provides equal scaling of all models that are trained in this study. In the dataset, the class distribution is moderate, although the non-diabetic samples prevail over the

DOI: 10.54254/2755-2721/2026.TJ31360

diabetic cases, an aspect that also applies to population-based screening cases, which is also being observed in long-term reviews of machine learning use in diabetes prediction [6].

2.2. Feature engineering

In attempt to estimate clinically significant nonlinear relationships that are not defined by the underlying features, this study proposed a set of interaction terms that are based on metabolic and demographic criteria. Particularly, the constructed interaction terms include Glucose x BMI, Glucose x Diabetes Pedigree Function, BMI x Insulin, Age x Pregnancies, Glucose x Age and Age x BMI. Such interactions represent composite physiological patterns, including obesity-glucose interaction, genetic predisposition to alteration of hereditary risks and age-reproductive stress interactions.

In order to avoid introducing noise due to an increased feature space, the evaluated interaction terms were considered using a dual-scoring process involving tree-based split gain and SHAP-based importance. The two measures were aggregated into a unified score by normalizing and achieving consolidation. Interaction features that were scored lowly were eliminated. The final feature representation of the interactions is a reduced 12-dimensional set of four clinically significant and high-contribution interaction terms and eight original predictors.

2.3. Model training

A set of standard machine learning models was trained in order to set some performance benchmarks against which comparative performance could be established in relation to machine learning-based diabetes prediction models, as described by Tasin et al. [7]. These are logistic regression, support vector machine, random forest, gradient boosting (XGBoost) and LightGBM. They combine linear decision functionality, kernel-based separations as well as a variety of nonlinear ensemble approaches. All models were trained with default or minimally tuned hyperparameters, which would guarantee a balanced comparison between the families of algorithms. Measurements of evaluation are accuracy, precision, recall, F1-score, and area under the ROC curve.

An ensemble stacking was also developed in order to investigate the question of whether heterogeneous learners can compensate each other through their predictive abilities. Random Forest, XGBoost, and LightGBM were both trained on the first layer, with the out-of-fold predictions forming part of the meta-features. The final predictions were then accomplished using the second-layer learner which was a logistic regression model. This was an integration of scaled probabilistic model of a linear classifier and expressive power of ensemble trees.

2.4. Final model and explainability

The last predictive model is the gradient-boosted tree classifier, for which the specific 12-feature representation has been chosen because gradient boosting-based algorithms have been shown to perform well when applied to diabetes classification tasks in previous research [8]. The values of hyperparameters (depth, the number of boosting rounds, sampling ratios, and learning rate) were selected so as to balance generalization and expressiveness. The performance was tested in three configurations, which were the baseline eight-feature model, the full interaction model with all constructed pairwise terms, and the reduced interaction-enhanced model. This comparison demonstrates the effect of interaction learning and importance-driven selection.

SHAP values were calculated using the final model to achieve a better understanding since SHAP has been extensively used to explain machine learning-based models of diabetes prediction because

DOI: 10.54254/2755-2721/2026.TJ31360

of its theoretical consistency and local accuracy [9]. Summary bar plots and beeswarm visualization were used to provide global explanations by highlighting prevalent predictors and general trends of variable influence. The dependence plots were also used to investigate nonlinear patterns and interaction effects captured by the model. Such explainability research offers clinically relevant information about the decision-making procedure of the model and justifies the validity of the chosen features, which is becoming more and more significant in clinically oriented and self-explainable diabetes prediction systems [10].

3. Results

3.1. Baseline model performance

The standardized eight-feature dataset was first used to test a set of baseline classifiers. Support vector machines and logistic regression achieved moderate performance, indicating the limited ability of both linear and margin-based decision functions to capture nonlinear metabolic relationships. Ensemble tree models, particularly XGBoost and LightGBM, showed significantly better predictive quality in terms of recall and AUC than linear baseline models. A visual analysis of the ROC curve is presented in Figure 1, which clearly demonstrates that the XGBoost curve is distinctly separated from the others and exhibits superior discriminative ability. Compared to all base learners, XGBoost produced the most balanced results in terms of accuracy, F1-score, and discriminative power. Accordingly, XGBoost was selected as the primary model for subsequent feature interaction analysis.

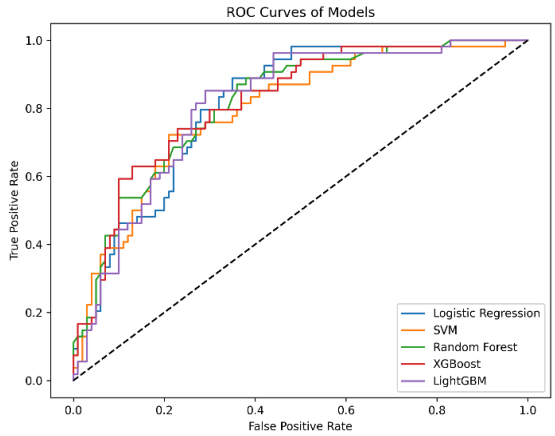

Figure 1. ROC curves of the baseline classifiers (picture credit: original)

3.2. Impact of feature interaction construction

To determine the effect of expanding the feature space with clinically informed interaction terms on predictive performance, XGBoost models were trained in two conditions, with the original eight-feature baseline and the complete interaction space of all six constructed pairwise terms. Recall and AUC did not improve systematically in response to the developments of interaction features, although both showed a small gain in overall accuracy. Specifically, the complete interaction model had indications of noise addition, as reflected by a slight decrease in AUC when compared with the eight features baseline even with a slight improvement in accuracy. These findings indicate that

DOI: 10.54254/2755-2721/2026.TJ31360

although metabolic interactions can boost the representational richness of the interaction, inclusion of all pairwise terms in a blind manner can impair the generalization.

3.3. Performance of the selected interaction feature set

The combined split-gain and SHAP importance criteria were applied and four out of the six interaction terms were retained and created the improved feature set to be used in the final model. This minimized, clinically significant, feature set produced the stable improvements in several evaluation measures as compared to the baseline as well as the full interaction features. The chosen-interaction model has better recall consistently and competitive AUC as shown in Figure 2, with the full interaction model demonstrating a reduced stability. In contrast to the full interaction model, the reduced feature set did not exhibit overfitting or performance instability which proved the effectiveness of guided feature pruning.

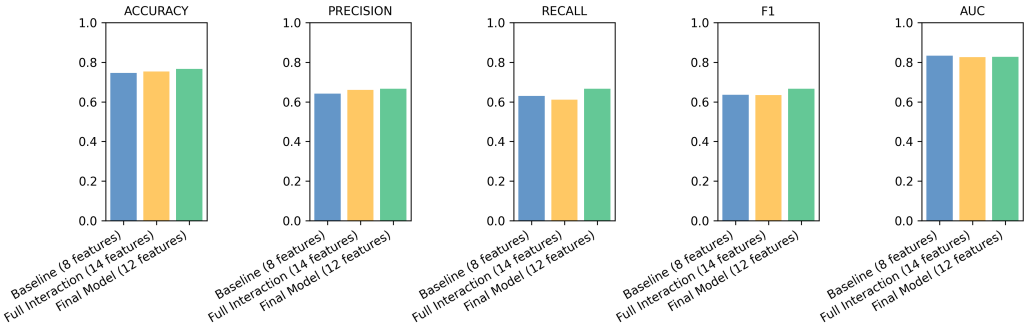

Figure 2. Performance comparison of baseline, full-interaction, and final models (picture credit: original)

In general, the comparison among the three configurations shows that predictive performance can be improved when interaction features are carefully selected, but the inclusion of all interactions built automatically can add noise and lead to degraded generalization in the models. Further, the combination of clinical priors with empirical importance scoring allows deriving a compact and strong feature representation, which enhances predictive sensitivity, a particularly important consideration when risk identification is performed in a medical context.

3.4. Stacking ensemble evaluation

A stacking structure was tested to understand whether the predictive performance can be further increased by using model aggregation. The ensemble performed better than the traditional linear models and was nearly on the level of the gradient boosting performance, although it did not always surpass the optimized interaction-based XGBoost model. This implies that, although heterogeneous learners extend the set of decision boundaries available, the gradient-boosted model, particularly with specialized interaction functionalities added, would already capture the majority of relevant patterns within the dataset. The stacking strategy therefore provided a favorable discriminatory cross-reference point. However, it was not chosen as the predominant forecasting platform.

DOI: 10.54254/2755-2721/2026.TJ31360

3.5. Ablation study

The contributions of individual features were further analyzed by systematic ablation analysis, in which each of eight original predictors was removed individually in an XGBoost classifier. As illustrated in Figure 3, removing glucose resulted in the largest decrease in AUC, indicating its dominant role in diabetes risk prediction. The next most pronounced performance drops were observed for BMI and age, suggesting that these variables also provide substantial discriminative information. In contrast, the removal of features such as insulin, skinfold thickness, and diabetes pedigree function led to only marginal changes in AUC, implying relatively limited contributions to the overall predictive performance.

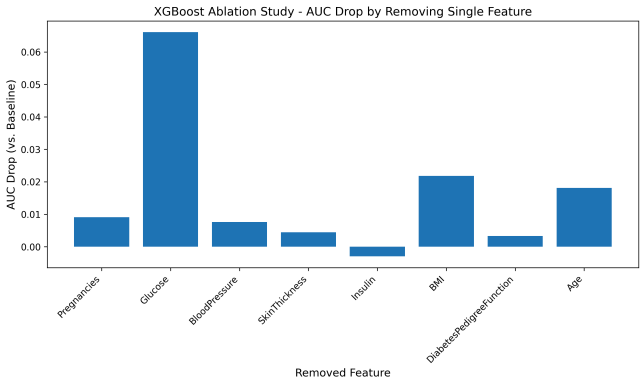

Figure 3. AUC drop after removing each original feature (picture credit: original)

3.6. SHAP-based model interpretation

To provide transparent interpretation of the final 12-feature model, SHAP values were computed across all test samples. Global summary plots identified glucose, BMI, age, and the interaction between glucose and BMI as dominant contributors to model predictions, consistent with established clinical knowledge regarding metabolic dysregulation and diabetes risk. Figure 4 presents the SHAP summary plot, illustrating the global impact of each predictor and the distribution of their contributions. The beeswarm visualization further revealed clear monotonic patterns for glucose and BMI as well as more nuanced nonlinear relationships for insulin and pedigree function. In particular, features with higher values, such as elevated glucose and BMI, shift predictions toward greater risk.

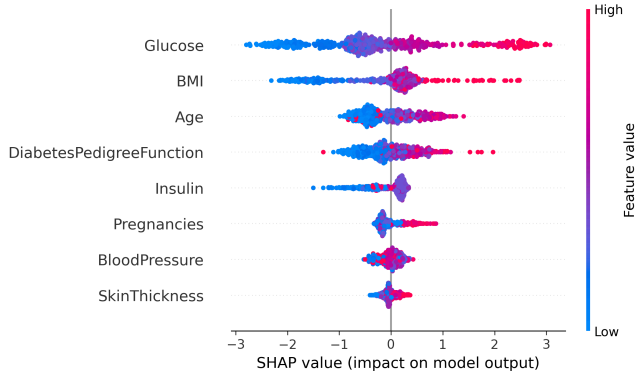

Figure 4. SHAP summary beeswarm plot of the final model (picture credit: original)

DOI: 10.54254/2755-2721/2026.TJ31360

Dependence plots revealed significant interactions implicitly captured by the model. To give an example, the contribution of BMI to the predicted risk increased with higher glucose levels, which validated the physiological relevance of the constructed interaction terms. In the same manner, the interplay between age and pregnancies indicated the metabolic load of the reproductive history. The SHAP-based insights provided above indicate that the model not only has a high predictive power but also provides clinically coherent interpretations, which suggests that this model can be applicable to medical decision-support scenarios.

4. Conclusion

The study proposed an interaction-enriched machine learning model to predict the risks of diabetes using structured clinical data. The model was compared against different baseline models, and feature interaction terms were introduced and refined in a more clinically oriented manner based on an importance-directed selection strategy combining the split gain and SHAP measures.

The XGBoost model fitted using the chosen set of interaction enhanced features had better recall scores and stable discriminatory performance than the baseline and full-interaction model.

Experimental evidence has shown that not all interactions are beneficial, and in certain cases, the blind addition of all interaction terms can lead to noise and adversely affect generalization, whereas the selective retention of clinically meaningful interactions can be used to make the model capture clinically significant nonlinear correlations that do not necessarily require any complexity increase. The analysis of feature ablation also validated glucose as the most significant predictor, followed by BMI and age, and supported the clinical plausibility of the trained hierarchy of features. The final model exhibited an attractive mix of sensitivity and discrimination when compared to the baseline models and the stacking ensemble, indicating that specialized interaction learning could be more effective than further model aggregation.

Moreover, SHAP-based interpretability presented clear and clinically coherent explanations of model behavior, which shows the worth of explainable artificial intelligence in medical risk assessment and justifies the reliability of the presented framework.

Despite these strengths, this study has a number of limitations. The experiments were conducted on one structured data set and generalization of the results should be further established in larger and more diverse populations. The future of such work might include the inclusion of other clinical / lifestyle variables, and prospective comparison in real-world screening scenarios. The extensions may also lead to the development of stronger and more useful interactive and interpretable diabetes prediction models.

References

[1] Rastogi, R. and Bansal, M. (2023) Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605. https: //doi.org/10.1016/j.measen.2022.100605

[2] Khanam, J.J. and Foo, S.Y. (2021) A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7, 432-439. https: //doi.org/10.1016/j.icte.2021.02.004

[3] Jaiswal, V., Negi, A. and Pal, T. (2021) A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), 435–443. https: //doi.org/10.1016/j.pcd.2021.02.005

[4] Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G. and Facchinetti, A. (2023) The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP. Scientific Reports, 13, Article 16865.

[5] Kee, O.T., Harun, H., Mustafa, N., Abdul Murad, N.A., Chin, S.F., Jaafar, R. and Abdullah, N. (2023) Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. Cardiovascular Diabetology, 22, Article 13.

DOI: 10.54254/2755-2721/2026.TJ31360

[6] Kiran, M., Xie, Y., Anjum, N., Ball, G., Pierscionek, B. and Russell, D. (2025) Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. Frontiers in Digital Health, 7, Article 1557467.
[7] Tasin, I., Nabil, T.U., Islam, S. and Khan, R. (2022) Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters. https: //doi.org/10.1049/htl2.12039
[8] Liu, Q., Ma, Y. and Cai, Y. (2023) Diabetes classification and prediction based on the XGBoost algorithm and its applications. Journal of Shanghai University of Medicine & Health Sciences, 42(4), 1–7.
[9] Ahmed, S., Kaiser, M.S., Hossain, M.S. and Andersson, K. (2024) A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. IEEE Access, 13, 37370–37388.
[10] Dharmarathne, G., Jayasinghe, T.N., Bogahawaththa, M., Meddage, D.P.P. and Rathnayake, U. (2024) A novel machine learning approach for diagnosing diabetes with a self-explainable interface. Healthcare Analytics, 5, 100301.