Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA36240

Comparing LightGBM and Deep Neural Networks for Early Diabetes Prediction

Xueyu Zhang[1*], Ziyan Zhao[2]

[1]Collegeof Science, Nanjing Forestry University, Nanjing, China
[2]School of Mathematics, Xi'an University of Finance and Economics, Xi'an, China
**Corresponding Author. Email: zxy180@njfu.edu.cn*

Abstract. Currently, diabetes affects around 589 million adults worldwide and is one of the more pressing chronic disease burdens in modern public health. Machine learning is widely accepted in diabetes risk modeling, although most published work tends to evaluate models in isolation rather than using them side by side in matching conditions. In particular, how LightGBM is compared to simple deep neural network structures in small, structured clinical datasets is not well described. This study directly addressed this question using a publicly available Kaggle dataset of 1,879 patient records and 46 clinical variables. The paper evaluated two models from three dimensions: differentiation, calibration, and interpretability. LightGBM is better overall than the DNN --AUC 0.965 vs. 0.913, accuracy 0.941. 0.870 – resulting in a significant reduction in missed diabetes cases. To explore what the model actually relies on, it applied SHAP analysis; fasting blood glucose and HbA1c were key drivers, consistent with clinicians' expectations. Overall, the results suggest that LightGBM is a more dependable option in situations where data are limited and interpretability is important, at least for this type of structured tabular input.

Keywords: LightGBM, Deep Neural Network, Diabetes Prediction, SHAP, Tabular Data

1. Introduction

Diabetes mellitus has emerged as one of the more burdensome chronic conditions of the past several decades, placing sustained pressure on healthcare systems across income levels and regions. The International Diabetes Federation estimates that roughly 589 million adults were living with diabetes as of 2024 — a figure expected to climb to around 853 million by 2050 [1]. Hyperglycemia can cause serious complications, affecting the heart, kidneys, eyes and nerves. Since there are often no early symptoms, early detection of high-risk groups is key to timely intervention.

Machine learning and deep learning approaches have been widely used in the early prediction of diabetes. Traditional algorithms such as support vector machines and decision trees have good classification performance on structured clinical datasets, while deep neural networks (DNNs) have good competition in various clinical prediction tasks due to their ability to automatically model nonlinear feature relationships [2-4]. However, these methods often face the challenge of unstable generalization performance when applied to small-scale tabular data. As Li et al. show, most standard machine learning algorithms struggle to meet performance thresholds in small-sample

© 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.BA36240

imbalanced clinical datasets, even when adjusted for high parameters [5]. At the same time, DNNs are increasingly used in clinical prediction tasks due to their modeling ability for complex nonlinear feature interactions. Grinsztajn et al., however, demonstrated through large-scale benchmarking of 45 tabular datasets that tree-based models consistently outperformed DNNs in small-scale and medium-sized structured data, attributing this gap to the sensitivity of DNNs to uninformative features and their inability to maintain the marginal distribution of initial tabular inputs [6]. This finding suggests that the use of DNNs in clinically listed settings warrants careful empirical validation, rather than hypothetical superiority.

Against this backdrop, Light Gradient Boosting Machine (LightGBM) has gained considerable attention in small- to medium-scale clinical data modeling due to its favorable balance between high predictive performance and low computational cost [7]. Wang et al. proposed a diabetes prediction framework integrating LightGBM with SHAP, demonstrating that LightGBM outperformed XGBoost and Random Forest on the Pima dataset, while SHAP analysis improved model interpretability by quantifying the contribution of individual clinical features to prediction outcomes [8]. Nevertheless, the comparison of the advantages of LightGBM over DNN-based approaches in terms of consistency across different structured clinical datasets, and accuracy-- interpretable trade-offs-- between the two model families is still insufficient. This study addresses this gap by directly experimentally comparing LightGBM and DNN in an independent small-sample clinical dataset, and applies SHAP-based interpretation to better-performing models to elucidate the relative importance of clinical predictors.

Addressing this discrepancy, the study employs an accessible Kaggle diabetes dataset (ages 20-90) to construct and assess a LightGBM model versus a basic DNN, analyzing it via classification metrics, confusion matrices, ROC curves, feature significance, and SHAP analysis. This comparison aims to provide practical guidance for model selection on small-sample clinical tabular data.

2. Methods

2.1. Dataset description

The dataset used in this study was obtained from the public repository Kaggle [9]. It is an open-access health-related dataset developed for diabetes prediction, containing demographic, lifestyle, medical history, clinical laboratory, and symptom variables. The original dataset consists of 1,879 records and 46 variables. As the data collection period is not disclosed, the dataset is regarded as cross-sectional, with each record representing an individual participant.

Predictor variables span three broad domains: patient background (age, gender), daily habits and behaviors (smoking status, alcohol use, exercise frequency, dietary patterns), and pre-existing health conditions (hypertension, family diabetes history), clinical and laboratory measurements (e.g., BMI, fasting blood sugar, HbA1c, and total cholesterol), and diabetes-related symptoms (e.g., frequent urination and excessive thirst). The target variable is Diagnosis, where 0 denotes non-diabetic individuals, and 1 denotes diabetic individuals.

Before model development, missing values and duplicate records were examined, and no missing or duplicate data were identified. Non-informative identifier variables, including Patient ID and attending physician's identifier, were removed, and data types were checked for consistency. All candidate predictors were retained for model construction, while feature importance and correlation analyses may be conducted in subsequent experiments for feature evaluation.

DOI: 10.54254/2755-2721/2026.BA36240

2.2. Method introduction

Two models — LightGBM and a DNN — were trained on the same dataset and evaluated under identical conditions, with SHAP applied afterward to examine what each model was relying on. To keep the class balance intact across both subsets, the data were divided into training and test portions using a stratified 80/20 split. After preprocessing, both LightGBM and DNN models were trained independently, followed by model evaluation, visualization, and SHAP-based feature contribution analysis.

LightGBM belongs to the gradient boosting family, where an ensemble of decision trees is built sequentially, each one correcting the residual errors left by its predecessors. Given a training set $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the feature vector of the i -th sample and y_i denotes the corresponding label, the final prediction is expressed as:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \tag{1}$$

where T is the total number of trees and f_t denotes the t -th tree. For each leaf node j, the optimal leaf weight w_j^* is computed as:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{2}$$

where I_j is the set of samples falling into leaf j, g_i and h_i are the first- and second-order gradients of the loss function with respect to the prediction of sample i, and λ is the L2 regularization coefficient. The best split is selected by maximizing the following gain:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{3}$$

where G_L, G_R , H_L , and H_R denote the sums of first- and second-order gradients for the left and right child nodes respectively, and γ is the regularization penalty on the number of leaf nodes.

The DNN architecture consists of an input layer, three hidden layers with 128, 64, and 32 neurons respectively, each applying the ReLU activation function, and an output layer with a sigmoid activation function for binary classification. The model is trained to minimize the binary cross-entropy loss:

$$\text{loss} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i log\left(p_i\right) + (1-y_i) log\left(1-p_i\right)\right) \tag{4}$$

where p_i is the predicted probability for sample i and $y_i \in \{0,1\}$ is the true label. The optimizer employed is Adam with a learning rate of 1×10^{-3}. To mitigate overfitting, dropout layers with a rate of 0.3 are applied after the first two hidden layers, and early stopping is employed with a patience of 15 epochs based on the AUC monitored on an internal validation set comprising 0.2 of the training data. The model is configured with a batch size of 32 and a maximum of 100 epochs, with the best weights restored upon early stopping.

For interpretability, SHAP was used to examine what the LightGBM model had learned. The test data X_{test} were first passed through the same preprocessing pipeline used during training — median imputation for continuous variables, one-hot encoding for categoricals — to keep the feature space

DOI: 10.54254/2755-2721/2026.BA36240

consistent. A TreeExplainer was then instantiated on the fitted model, and SHAP values were computed for the held-out test set.

The primary visualization was a SHAP summary plot, where features are ordered top to bottom by how much they moved predictions on average, and the color of each dot reflects whether a given feature value pushed the output higher or lower.

3. Results and analysis

3.1. Model comparison

Table 1 shows how the two models performed on the test set. LightGBM came out ahead on every metric — AUC, accuracy, precision, recall, and F1. The gaps were most obvious in AUC, accuracy, and F1. In other words, LightGBM was better both at separating diabetic from non-diabetic cases and at balancing precision against recall.

Table 1. Performance comparison of LightGBM and DNN models on the test set

Model	AUC	Accuracy	Precision	Recall	F1
LightGBM	0.965	0.941	0.951	0.900	0.925
DNN	0.913	0.870	0.848	0.820	0.834

3.2. ROC curve analysis

Figure 1 plots the ROC curves for both models. LightGBM's curve sits closer to the upper-left corner throughout, which lines up with the numbers in Table 3. One thing worth pointing out: at a false positive rate below 0.05, LightGBM's true positive rate still exceeded 0.90 — meaning it held onto strong sensitivity without sacrificing much specificity. That kind of behavior is useful if the goal is early screening, where catching real cases matters more than avoiding every false alarm.

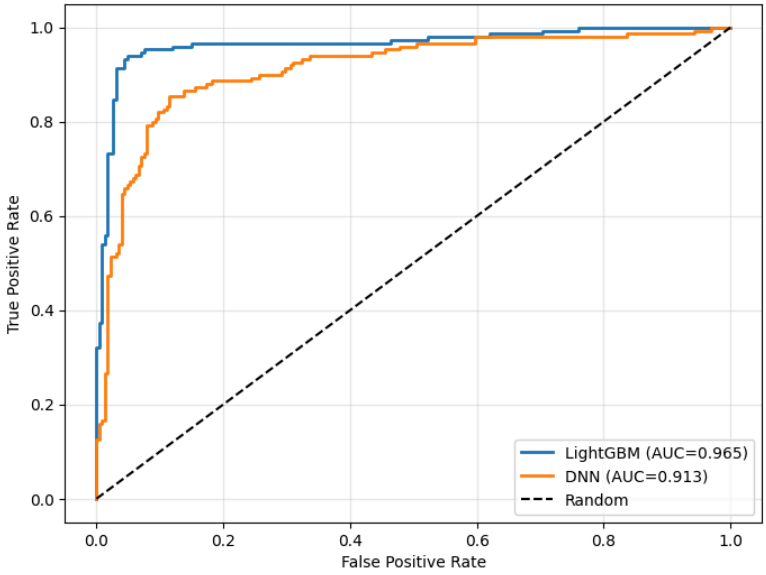

Figure 1. ROC curve comparison (photo/picture credit: original)

DOI: 10.54254/2755-2721/2026.BA36240

To further illustrate the class-specific prediction performance of the two models, Figure 2 breaks things down further with confusion matrices.

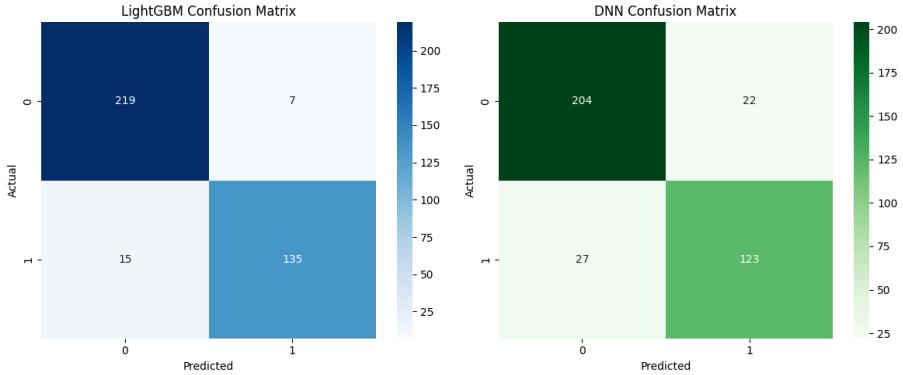

Figure 2. Confusion matrices of the LightGBM and DNN models (photo/picture credit: original)

LightGBM got 354 out of 376 test samples right. Of the 226 non-diabetic cases, 219 were correctly identified, yielding a specificity of 0.969. For the 150 diabetic subjects (Class 1), 135 were correctly detected, corresponding to a recall of 0.900. Only 7 false positives and 15 false negatives were observed, indicating a relatively low misclassification rate.

In contrast, the DNN model produced 22 false positives and 27 false negatives, both substantially higher than those of LightGBM. More missed diabetic cases are the bigger concern clinically, since those are the patients who leave without a diagnosis.

3.3. Feature importance and SHAP interpretation

Figure 3 shows the top 20 features ranked by LightGBM's split-count importance. Fasting plasma glucose and HbA1c scored highest at around 880 and 820, respectively. Age, fatigue levels, and physical activity rounded out the top five. Health literacy, cholesterol triglycerides, frequent urination, HDL-C, and systolic blood pressure also made the top 20.

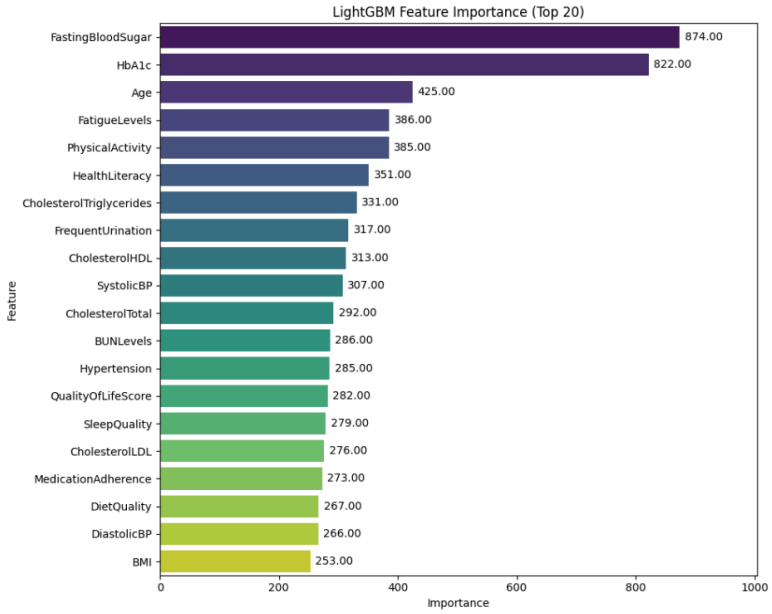

Figure 3. LightGBM feature importance (top20) (photo/picture credit: original)

DOI: 10.54254/2755-2721/2026.BA36240

Since split-based feature importance only reflects how frequently a feature is used during tree construction, SHAP analysis was further performed to evaluate the direction and magnitude of feature contributions.

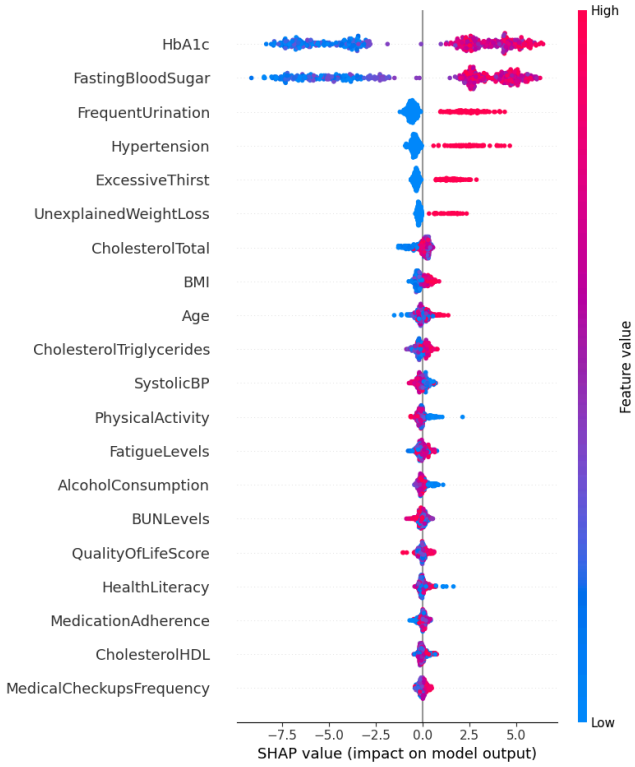

Figure 4. SHAP value (impact on model output) (photo/picture credit: original)

Figure 4 shows the SHAP summary plot of the LightGBM model. HbA1c and fasting blood glucose had the widest SHAP distributions, meaning they pushed predictions the most. Higher values of these features were generally associated with positive SHAP values, whereas lower values tended to produce negative SHAP values. FrequentUrination, Hypertension, ExcessiveThirst, and UnexplainedWeightLoss also contributed, though less dramatically.

It is worth noting that the feature importance ranking derived from LightGBM's built-in split-count metric differs from the SHAP-based ranking. For instance, FatigueLevels ranks prominently in Figure 4 but shows comparatively lower SHAP values. This happens because split-count reflects how often a feature gets chosen by the tree-building algorithm — which is partly just an artifact of how that feature's values are distributed — whereas SHAP measures the actual marginal contribution to each individual prediction, with feature interactions accounted for. The SHAP ranking is treated as primary here for that reason.

3.4. Discussion

LightGBM's AUC of 0.965 fits with what the broader literature has shown for gradient boosting on structured clinical data. Rufo et al. reported an AUC of 0.981 with LightGBM on a real-world diabetes dataset, outpacing KNN, SVM, Naive Bayes, Bagging, Random Forest, and XGBoost [10]. The present study isn't trying to match those numbers directly — different dataset sizes, feature sets, and populations make direct comparison tricky — but it does add evidence that LightGBM holds up

DOI: 10.54254/2755-2721/2026.BA36240

reasonably well even on small synthetic tabular data, where DNNs tend to struggle without large sample volumes to learn from [6].

From a clinical perspective, LightGBM generated fewer false negatives than DNN (15 vs. 27), reducing the likelihood of missed diabetes cases. Given that undiagnosed diabetes significantly increases the risk of severe complications and escalating healthcare costs, the lower false-negative rate of LightGBM may be particularly valuable for early screening in primary healthcare settings [11].

A few limitations are worth being upfront about. The dataset is synthetic, built to resemble real patient data rather than actually coming from a clinic; the results are better understood as methodological rather than clinical evidence. There was also no external validation set, so how well either model generalizes to different populations is an open question. Real-world multicenter data and external cohorts would be the natural next step.

4. Conclusion

This study offers a methodological comparison of LightGBM and DNN for structured tabular prediction tasks by evaluating both models on the same synthetic clinical dataset under identical experimental conditions. The results indicate that under limited data conditions, the tree-based ensemble model achieved superior classification performance relative to the deep neural network across all evaluation metrics, which is consistent with the broader literature on the behavior of tree-based models on small-scale tabular data.

The SHAP results added something beyond the performance numbers. Fasting blood glucose, HbA1c, age, fatigue, and physical activity came out as the dominant predictors — which is essentially what a clinician would have guessed. That correspondence matters. A model that achieves high AUC but surfaces predictors with no clinical plausibility is hard to trust, regardless of what the test metrics say. The fact that LightGBM's internal logic tracks established diabetes risk knowledge makes its outputs easier to reason about and, arguably, easier to act on.

That said, the practical ceiling here is real. The dataset is synthetic, and one well-performing experiment on simulated data doesn't settle the question of clinical utility. Whether these results hold on actual patient records from multiple centers — with all the messiness that entails — remains to be tested. Until that kind of external validation exists, the findings are best read as evidence that this modeling approach is worth pursuing further, not as a deployment-ready solution.

Author contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] International Diabetes Federation. (2025). *IDF diabetes atlas* (11th ed.). International Diabetes Federation.
[2] Kumari, A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications, 3*(2), 1797–1801.
[3] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process, 5*(1), 1–14.
[4] Ashiquzzaman, A., Tushar, K., Islam, A. K., et al. (2017). Reduction of overfitting in diabetes prediction using deep learning neural network. In *IT Convergence and Security 2017* (pp. 35–43). Springer. https://doi.org/10.1007/978-981-10-6451-7_5
[5] Li, G., Li, C., Wang, C., & Wang, Z. (2024). Suboptimal capability of individual machine learning algorithms in modeling small-scale imbalanced clinical data of local hospital. *PLOS ONE, 19*(2), Article e0298328.

DOI: 10.54254/2755-2721/2026.BA36240

https://doi.org/10.1371/journal.pone.0298328
[6] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems, 35*, 507–520.
[7] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3149–3157.
[8] Wang, W., Liao, B., Li, M., & Sun, R. (2022). Combination of LightGBM and SHAP for diabetes prediction and feature analysis. *Journal of Chinese Computer Systems, 43*(9), 1877–1885.
[9] zyh1104. (2024). 027_Diabetes data [Data set]. *Kaggle*. https://www.kaggle.com/datasets/zyh1104/027-shujuji
[10] Rufo, D. D., Debelee, T. G., Ibenthal, A., et al. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics, 11*(9), Article 1714. https://doi.org/10.3390/diagnostics11091714
[11] American Diabetes Association Professional Practice Committee. (2025). 3. Prevention or delay of diabetes and associated comorbidities: Standards of Care in Diabetes—2025. *Diabetes Care, 48*(Suppl 1), S50–S58. https://doi.org/10.2337/dc25-S003