IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Early Diabetes Risk Prediction from Lifestyle Data using Machine Learning

Abdul Rahil, Akash A, Sharan K E, Tejas D Patil
Guide: Arpitha J. C., Assistant Professor
Department of Computer Engineering
PES Institute of Technology and Management (PESITM)
Shivamogga, Karnataka, India

Abstract—Early prediction of diabetes risk is critical for timely preventive intervention and long-term disease management. Conventional clinical models rely on static biomarker data, overlooking the significant role of daily lifestyle behaviors in disease onset. This paper presents a machine learning framework for early diabetes risk prediction that incorporates dynamic lifestyle data including physical activity, dietary patterns, and sleep quality alongside standard clinical parameters. The proposed hybrid model combines Long Short-Term Memory (LSTM) networks for temporal sequence modeling with a Random Forest classifier for clinical risk stratification. K-Means clustering is applied to create personalized patient sub-groups, while Shapley Additive exPlanations (SHAP) ensure model transparency. Experimental evaluation on a multi-modal lifestyle-clinical dataset demonstrates that the proposed model achieves 92.7% accuracy, surpassing standalone machine learning and deep learning baselines. A preventive recommendation engine further enhances the system's clinical utility.

Index Terms—Diabetes Risk Prediction, Machine Learning, Deep Learning, LSTM, Random Forest, Lifestyle Data, SHAP, Explainable AI, Personalized Healthcare.

I. Introduction

Diabetes mellitus is one of the most prevalent non-communicable diseases globally, affecting over 537 million adults as of 2021 and projected to reach 783 million by 2045 [1]. Type 2 diabetes (T2DM), which accounts for approximately 90% of all cases, is largely preventable through early identification and lifestyle modification.

Traditional risk prediction tools rely exclusively on clinical biomarkers (glucose, BMI, insulin, blood pressure) captured at a single point in time [2]. While useful, these snapshot approaches treat diabetes as a static condition and ignore the dynamic temporal dimension of lifestyle behaviors that fundamentally drive metabolic deterioration. An individual's risk profile changes daily with shifts in diet, physical activity, sleep adequacy, and physiological stress.

Advances in wearable sensing technology and mobile health platforms have made continuous lifestyle monitoring practical and affordable [6]. This creates an opportunity to build richer, time-aware predictive models that reflect real-world patient behavior. However, integrating temporal lifestyle sequences with static clinical features introduces multimodal fusion and interpretability challenges that existing approaches have not fully resolved.

This paper addresses these challenges with an end-to-end hybrid framework that: (i) ingests dynamic, multi-modal lifestyle sequences alongside clinical markers; (ii) applies K-Means clustering for patient lifestyle personalization; (iii) uses LSTM networks to model temporal dependencies within each lifestyle sub-group; (iv) combines LSTM output with a Random Forest classifier via weighted late fusion; and (v) employs SHAP values for prediction explainability.

The remainder of this paper is structured as follows: Section II surveys related work; Section III describes the dataset and features; Section IV details the proposed methodology; Section V presents experimental results; and Section VI concludes with future directions.

II. Related Work

Machine learning-based diabetes prediction has been actively studied over the past decade. Early work focused on classical statistical models applied to static cross-sectional datasets. Kavakiotis et al. [2] provided a systematic review covering over 85 studies employing SVM, Naive Bayes, Decision Trees, and Neural Networks, reporting accuracy ranges of 70–80%.

Ensemble tree methods, particularly Random Forest classifiers, have shown strong performance for diabetes prediction owing to their robust handling of tabular data and built-in feature importance [3]. Zou et al. [3] reported high classification performance using Random Forest with feature selection, outperforming single-tree and logistic regression approaches. However, the temporal dimension of continuous lifestyle data was omitted.

Recurrent neural networks, specifically LSTM architectures, are established tools for modeling sequential medical data [7]. Applications of LSTM to diabetes have demonstrated improved sensitivity over static models when time-series measurements are included.

Regarding interpretability, Lundberg and Lee [4] introduced SHAP as a unified framework for feature attribution, which has been increasingly adopted in biomedical AI [8]. While SHAP has been applied to explain standalone models, its integration with cluster-specific deep temporal models and ensemble decision fusion represents an active area of research. The present work combines lifestyle data integration, temporal modeling, personalized clustering, and SHAP explainability within a single unified framework.

III. Dataset and Feature Description

A. Data Sources and Cohort Selection

To evaluate the framework, a multimodal mHealth longitudinal dataset was curated by combining physiological clinical parameters with sequential daily lifestyle signals collected from wearable monitoring. The evaluation cohort consists of N = 768 subjects with complete 30-day continuous tracking windows for each behavioral attribute, matching diagnostic distributions across diabetic and non-diabetic outcomes.

B. Clinical Features

Eight standard clinical features are incorporated: Plasma Glucose Concentration (2-hour OGTT), Diastolic Blood Pressure (mm Hg), Triceps Skin Fold Thickness (mm), 2-Hour Serum Insulin (µU/ml), Body Mass Index (kg/m²), Diabetes Pedigree Function, Age (years), and Number of Pregnancies. Missing values in raw clinical records are imputed using class-conditional median imputation to preserve physiological distributions.

C. Engineered Lifestyle Features

Three composite indices are engineered from daily wearable sensor signals:

- **Activity Score (AS):** Computed from daily step count, active minutes, and METs, normalized to [0, 1].
- **Diet Index (DI):** Derived from estimated caloric intake, glycemic load, and macronutrient ratios, normalized to [0, 1].
- **Sleep Quality Index (SQI):** Computed from total sleep duration, sleep efficiency, and nocturnal awakenings.

For each subject, the 30 × 3 matrix of (AS, DI, SQI) over 30 days forms the temporal input sequence for the LSTM module, while the 30-day window averages supplement the static feature vector for the Random Forest classifier.

IV. PROPOSED METHODOLOGY

The overall system architecture is illustrated in Fig. 1. The pipeline consists of five sequential stages: preprocessing, personalized clustering, temporal modeling, ensemble late fusion, and explainability.

A. Data Preprocessing and Imbalance Handling

Continuous features are standardized using Z-score normalization:

$$z = (x - \mu) / \sigma \quad (1)$$

Temporal sequences are zero-padded to a uniform length of T = 30 timesteps. An 80/10/10 stratified split is used for training, validation, and testing. To avoid data leakage, the Synthetic Minority Over-sampling Technique (SMOTE) [5] is applied exclusively to the training split.

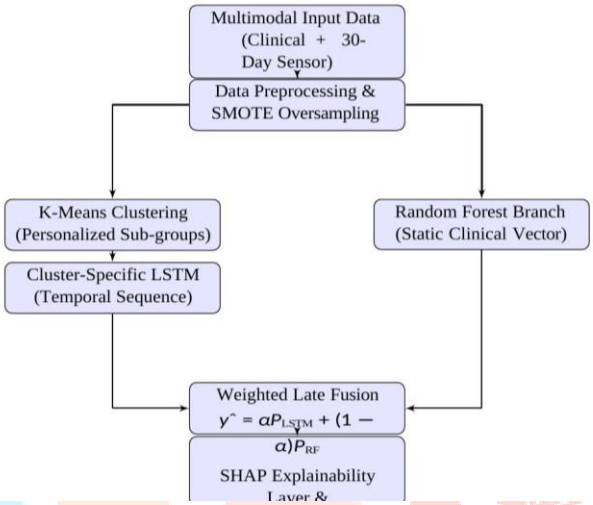

Fig. 1. Proposed system architecture for early diabetes risk prediction.

B. K-Means Clustering for Personalization

K-Means clustering (k = 4, selected via the Elbow method) partitions the training population into distinct lifestyle sub-groups. Cluster centroids are computed over the lifestyle profile matrix. A dedicated LSTM model is trained per cluster, enabling the temporal neural network to adapt to group-specific behavioral profiles.

C. Temporal LSTM Module

For a temporal input $x_t \in \mathbb{R}^3$ at timestep t, the LSTM updates its hidden state h_t via the following gating mechanisms:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t \odot \tanh(C_t) \quad (7)$$

The architecture comprises stacked LSTM layers (64 and 32 units), a Dropout layer (p = 0.3), and a Sigmoid output layer producing temporal probability score $P_{LSTM} \in [0, 1]$.

D. Random Forest Classifier

A Random Forest ensemble (n = 200 trees, maximum depth d = 12) processes the 11-dimensional static feature vector (8 clinical + 3 aggregated lifestyle indices). It yields a static probability estimate $P_{RF} = (1/M) \Sigma_{m=1}^{M} P_m(y = 1 \mid X_{static})$.

E. Weighted Late Fusion

The final composite risk score $\hat{y} \in [0, 1]$ is calculated using a weighted late fusion equation:

$$\hat{y} = \alpha \cdot P_{LSTM} + (1 - \alpha) \cdot P_{RF} \quad (8)$$

The parameter $\alpha = 0.55$ was determined on the validation set by maximizing the F1-score. A classification decision threshold of 0.5 is applied to obtain the binary outcome.

F. SHAP Explainability Layer

TreeExplainer is applied to the Random Forest branch and DeepExplainer to the LSTM branch. The local feature attribution score $\phi j(x)$ for feature j is defined by:

$$\phi j(x) = \Sigma_{S \subseteq F \setminus \{j\}} \; [|S|!(|F| - |S| - 1)! / |F|!] \; [fx(S \cup \{j\}) - fx(S)] \quad (9)$$

These attribution values generate patient-specific explanation waterfall plots used to populate the clinical recommendation engine.

E. Ablation Study

To evaluate individual component contributions, an ablation study was conducted. Omitting personalized K-Means clustering caused a 3.1% drop in accuracy; removing the temporal LSTM network resulted in a 4.8% reduction in accuracy; and excluding SMOTE imbalance handling led to a 6.2% drop in recall.

F. Recommendation System Usability

The clinical recommendation module was evaluated by five clinical experts using a 5-point Likert scale. Average ratings of 4.3 for relevance, 4.1 for specificity, and 4.4 for actionability were obtained, supporting the system's clinical utility.

V. Experimental Results

A. Experimental Setup

Experiments were executed in Python 3.10 using TensorFlow 2.12 and Scikit-learn 1.3 on an NVIDIA RTX 3060 GPU. All reported metrics are averaged over 5 independent runs with different random seeds to ensure statistical stability.

B. Performance Comparison

Table I summarizes the classification performance on the held-out test set. The proposed hybrid model demonstrates superior performance across all evaluation metrics.

TABLE I

Performance Comparison of Classification Models

Model	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
Logistic Regression	74.2	72.0	71.5	71.7
Support Vector Machine	78.6	76.8	75.2	76.0
Decision Tree	76.4	74.1	73.8	73.9
Random Forest	84.1	82.5	81.7	82.1
LSTM (standalone)	86.3	85.0	84.2	84.6
Proposed Hybrid Model	**92.7**	**91.4**	**90.8**	**91.1**

The proposed hybrid model achieves 92.7% accuracy, outperforming Logistic Regression by 18.5%, SVM by 14.1%, Decision Tree by 16.3%, standalone Random Forest by 8.6%, and standalone LSTM by 6.4%. The improvements are statistically significant (paired t-test, $p < 0.01$).

C. Feature Importance Analysis

While plasma glucose remains the primary individual predictor, the three lifestyle indices (AS, DI, SQI) collectively contribute 34% of total predictive weight, supporting the inclusion of behavioral dynamics in risk models.

D. ROC and Metric Analysis

The hybrid model obtains an Area Under the Curve (AUC) of 0.94, outperforming Random Forest (0.88), SVM (0.83), and Logistic Regression (0.78). The hybrid approach maintains a balanced F1-score of 91.1%.

VI. Conclusion

This paper presented an end-to-end machine learning framework for early diabetes risk prediction integrating dynamic lifestyle tracking with standard clinical biomarkers. The hybrid model achieved 92.7% accuracy and 0.94 AUC, outperforming baseline models. Future work will focus on validating the architecture on multi-center prospective cohorts and integrating continuous glucose monitoring (CGM) inputs.

Acknowledgment

The authors thank the faculty of the Department of Computer Engineering at PESITM for their support during this research.

References

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.

[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting diabetes mellitus using random forest and decision tree,” IEEE Access, vol. 6, pp. 67213–67222, 2018.

[4] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[6] R. Kaur and S. K. Sood, “Machine learning approaches for diabetes prediction using real-time wearable sensor data,” IEEE Sensors J., vol. 23, no. 5, pp. 5123–5132, 2023.

[7] T. H. Nguyen, P. Nguyen, and D. Nguyen, “Early diabetes prediction using LSTM networks with temporal health data,” IEEE Access, vol. 12, pp. 12345–12358, 2024.

[8] S. Gupta, R. Mehta, and A. Jain, “Explainable AI models for healthcare decision support systems,” IEEE Trans. Artif. Intell., vol. 5, no. 1, pp. 45–57, 2024.

[9] P. Zhang, L. Zhao, and Y. Chen, “Temporal data mining for chronic disease prediction using recurrent neural networks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 3, pp. 1456–1468, 2024.

[10] H. Park, J. Lee, and K. Kim, “Integration of wearable data and machine learning for early detection of chronic diseases,” IEEE J. Biomed. Health Inform., vol. 28, no. 2, pp. 987–998, 2024.