Binary Classification of Type 2 Diabetes Using Wearable Activity, Continuous Glucose Monitor, and Clinical Record Data

AI-Researcher

Abstract

Diabetes has emerged as a critical global health issue, with millions affected worldwide and a sharp increase in prevalence anticipated. Traditional methods of diabetes classification primarily depend on static clinical data, which often overlook the dynamic aspects of diabetes management enabled by continuous monitoring technologies, leading to delays in accurate diagnosis and ineffective individualized health management. To tackle these challenges, we propose a novel binary classification methodology that effectively distinguishes diabetic individuals from non-diabetics through an integrated feature set sourced from multimodal biomedical data. Our approach emphasizes a systematic feature extraction process, utilizing data from wearable technologies, Continuous Glucose Monitoring (CGM) systems, and structured clinical data while ensuring that label leakage is prevented. Experimental results demonstrate significant improvements in classification accuracy, as indicated by metrics such as AUROC and Macro F1 scores. This research contributes to enhanced predictive capabilities in diabetes classification and serves as a foundation for more tailored diabetes management strategies, ultimately aiming to improve patient care and health outcomes.

1 Introduction

Diabetes is an increasingly prevalent global health concern, leading to significant challenges for healthcare systems worldwide. The International Diabetes Federation reports that approximately 463 million adults were living with diabetes in 2019, a figure projected to rise substantially in the coming decades [12]. Effective early detection and classification of diabetes are crucial for implementing timely interventions, improving patient outcomes, and minimizing long-term complications associated with the disease. Traditional diabetes classification methods have largely depended on clinical data, which often fail to capture the dynamic nuances afforded by continuous monitoring technologies. Limitations in current methodologies may impede timely diagnosis, particularly of diabetes subtypes, resulting in suboptimal health management for affected individuals.

Recent advancements in machine learning and wearable technology have fostered interest in multimodal approaches to diabetes classification. Researchers have begun investigating various algorithms and data sources, including wearable activity monitors and Continuous Glucose Monitoring (CGM) systems, to enhance prediction accuracy and model robustness [14]. These studies underscore the importance of feature extraction from diverse data modalities, demonstrating that comprehensive datasets significantly improve classifier performance [15]. Nonetheless, challenges remain, particularly in differentiating between complex diabetes subtypes while mitigating the risk of label leakage during data preprocessing [13]. This situation prompts essential research questions regarding the effectiveness of integrating multimodal data and its potential impact on classification accuracy.

In this context, we propose a novel methodology for binary classification aimed at distinguishing diabetic individuals from non-diabetic counterparts using a combined feature vector derived from multimodal biomedical data. Our approach adheres to a systematic data preprocessing framework that integrates physiological and behavioral features from wearable technology, glycemic variability metrics from CGM, and clinical data structured according to the OMOP Common Data Model, all while meticulously avoiding potential label leakage. This comprehensive framework addresses the challenges associated with data integration and feature extraction, ultimately enhancing predictive power in diabetes classification.

The contributions of this work can be summarized as follows:

- Development of an innovative methodological framework for diabetes classification that leverages multimodal data integration and robust feature extraction techniques.
- Comprehensive exploration of the interplay between wearable data, CGM metrics, and clinical data, providing insights into pathways for enhanced predictive accuracy.
- Empirical validation of the proposed methodology through rigorous evaluations utilizing established metrics such as AUROC and Macro F1 scores, demonstrating its efficacy compared to traditional approaches.
- Insights from this research will inform future investigations into personalized diabetes management and classification practices, ultimately contributing to improved patient care.

2 Related Work

2.1 Machine Learning Frameworks

Recent advancements in machine learning have been significantly influenced by versatile frameworks that provide robust tools for model development and evaluation. Scikit-learn, a prominent library in Python, has introduced modularity that allows users to implement a variety of algorithms with ease, which has led to its wide adoption in both academia and industry [1]. Furthermore, the comprehensive guide "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" elaborates on practical approaches for employing these frameworks, demonstrating their application across different machine learning tasks [7]. Other significant frameworks include TensorFlow and PyTorch, which have contributed to deep learning advancements by providing extensive flexibility and support for large-scale computations [4, 5]. However, existing frameworks often face challenges in terms of scalability, user-friendliness, and integration with emerging technologies. Our work builds upon these frameworks by enhancing their accessibility and scalability, particularly for users with varying levels of expertise.

2.2 Tree-Based Methods

Tree-based methods, particularly Random Forests and XGBoost, have established themselves as leading techniques in predictive modeling due to their high accuracy and efficiency. The Random Forest algorithm aggregates the predictions of multiple decision trees to improve overall performance and robustness against overfitting [3]. XGBoost, a variant of gradient boosting, has gained attention for its speed and performance, thanks to optimizations such as parallel computing and tree pruning [2]. Additional developments like LightGBM have further pushed the boundaries of scalability and speed in tree-based learning approaches [6]. Nevertheless, limitations such as interpretability and sensitivity to noise remain challenges for these methodologies. Our research emphasizes the interpretability aspect of tree-based models while maintaining their predictive performance, contributing to the ongoing discourse in this domain.

2.3 Educational Approaches in Machine Learning

Education in machine learning has evolved rapidly, with innovative approaches such as MOOCs playing a crucial role in knowledge dissemination. The course "Machine Learning in Python with Scikit-learn" exemplifies how structured online learning can enhance comprehension of machine learning principles and applications [8]. Additionally, hands-on tutorials and workshops are proliferating, providing real-world data sets and challenges that bridge theoretical knowledge and practical skills [10, 9]. Furthermore, the integration of interactive learning platforms has made machine learning more accessible to diverse audiences [11]. Despite these advancements, there remains a gap in resources that cater to personalized learning experiences. Our work aims to address this gap by proposing an adaptive educational model that tailors machine learning instruction to individual learning styles and paces.

3 Proposed Methodology for Diabetes Classification

This proposed methodology outlines a comprehensive framework for a binary classification task aimed at distinguishing diabetic individuals from non-diabetic individuals. The framework utilizes various machine learning algorithms, including Logistic Regression, Random Forest, Gradient Boosting (XGBoost), and Support Vector Machines (SVM), to enhance predictive accuracy and robustness. Furthermore, the approach integrates multi-modal biomedical data comprising features derived from wearable activity monitors, Continuous Glucose Monitoring (CGM) data, and clinical data structured according to the OMOP Common Data Model.

3.1 Data Transformation and Feature Normalization

Data preprocessing is an essential phase that prepares raw multi-modal input data for effective machine learning classification aimed at distinguishing diabetic from non-diabetic individuals. This phase addresses significant challenges, such as preventing label leakage and ensuring high-quality data. The objective is to create a normalized and integrated feature set that provides robust support for the training and validation of predictive models.

The data preprocessing workflow can be illustrated as follows:

$$\text{Raw Multi-Modal Data} \rightarrow \text{Data Preprocessing} \rightarrow \text{Normalized Features}$$

3.1.1 Physiological and Behavioral Feature Extraction from Wearables

This subcomponent emphasizes the extraction of physiological and behavioral features from raw signals of wearable activity monitors. The feature extraction process computes various statistical metrics over specified time windows or the entire duration of each signal, which enhances the classifiers' predictive accuracy. Key metrics computed include:

- Mean
- Standard Deviation
- Minimum
- Maximum
- Median
- Interquartile Range (IQR)
- Signal Magnitude Area (SMA)
- Zero-Crossing Rate (ZCR)

These selected features are correlated with physical activities relevant to diabetes risk assessment and management.

Input: Raw JSON signals from wearable activity monitors.
Output: A participant-specific statistical feature vector constructed from the aforementioned metrics.
Workflow:

$$\text{Raw JSON Signals} \rightarrow \text{Feature Extraction} \rightarrow \text{Participant Feature Vectors}$$

The extracted statistical features yield critical insights into participants' dietary habits and overall physical engagement, which are vital for the study's objectives.

3.1.2 Glycemic Variability Metrics Extraction from CGM Data

This section focuses on extracting significant metrics of glycemic variability from Continuous Glucose Monitoring (CGM) data, which are crucial for understanding individual glucose patterns in diabetes management. Key features calculated include:

- Estimated A1c (eA1c)
- Glucose Management Indicator (GMI)
- Mean Amplitude of Glycemic Excursion (MAGE)
- Continuous Overall Net Glycemic Action (CONGA)
- Mean of Daily Differences (MODD)
- Glycemic Indices (LBGI/HBGI)
- Time-in-Range (TIR), Time-Below-Range (TBR), and Time-Above-Range (TAR)
- Area Under the Curve (AUC)

To ensure the reliability of these features, CGM data undergoes rigorous preprocessing techniques, including noise filtering, interpolation of missing values, and outlier detection. The systematic execution of these steps ensures valid feature extraction for subsequent modeling tasks.

Input: Processed CGM time-series data.
Output: A participant-specific glycemic variability feature vector.
Workflow:

$$\text{CGM Time Series} \rightarrow \text{Feature Extraction} \rightarrow \text{Glycemic Variability Features}$$

This meticulous extraction contributes invaluable insights into glucose management and variability, which is fundamental for predicting diabetes-related outcomes.

3.1.3 Clinical Feature Aggregation from OMOP Data

The final preprocessing component aggregates clinical features extracted from OMOP Common Data Model (CDM) tables. During this aggregation, we ensure the exclusion of diabetes-related data to prevent label leakage. The process consolidates demographic information, counts of unique medical conditions, and various notable measurements captured throughout clinical interactions. Particularly, explicit diabetes indicators such as 'diabetes', 'dm2', 'predm', 'a1c', and 'hemoglobin A1c' are excluded to safeguard the integrity of the modeling task.

Input: Clinical data extracted from OMOP tables.
Output: An aggregated clinical feature vector for each participant.
Workflow:

$$\text{OMOP Clinical Data} \rightarrow \text{Feature Aggregation} \rightarrow \text{Clinical Feature Vectors}$$

Through this systematic aggregation, participants' overall health context is captured without introducing direct outcome leakage, providing a critical foundation for predictive modeling efforts.

3.2 Classifier Training and Evaluation

This section delineates the training methodology for various baseline classifiers using the normalized and processed feature sets generated during the preprocessing phase. The primary objective is to independently train each classifier for binary classification—distinguishing diabetic individuals from non-diabetic participants. This approach allows for a comparative analysis of the classifiers, elucidating their strengths and weaknesses in relation to the multi-modal biomedical dataset used in this study.

3.2.1 Selected Classifiers for Diabetes Classification

In this analysis, we focus on four fundamental classifiers: Logistic Regression, Random Forest, Gradient Boosting (XGBoost), and Support Vector Machine (SVM). These classifiers are selected based on their distinctive modeling characteristics and robustness in handling intricate features present in biomedical data.

3.2.2 Classifier Model Architectures and Mechanisms

Each classifier is initialized with optimally tuned hyperparameters specific to our classification task. Below is a summary of the key characteristics and operational mechanics of each model:

- **Logistic Regression:** The model estimates class membership probabilities $P(y = 1|\mathbf{x})$ using the logistic function:
$$P(y = 1|\mathbf{x}; \mathbf{w}, b) = \sigma(\mathbf{w}^T\mathbf{x} + b)$$
To train, we minimize binary cross-entropy loss:
$$L(\mathbf{w}, b) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(P(y = 1|\mathbf{x}_i; \mathbf{w}, b)) + (1 - y_i)\log(1 - P(y = 1|\mathbf{x}_i; \mathbf{w}, b))]$$

- **Random Forest:** This ensemble method constructs multiple decision trees, aggregating predictions via majority voting:
$$H(\mathbf{x}) = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\}$$

- **Gradient Boosting (XGBoost):** It builds trees sequentially, aiming to minimize the logistic loss:
$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

- **Support Vector Machine (SVM):** This algorithm constructs a hyperplane maximizing the margin between classes:
$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$
For non-linear separation, kernel functions are utilized, leading to the optimization problem defined as:
$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$$

3.2.3 Model Standardization through Pipelines

All classifiers are integrated into a processing pipeline utilizing the `StandardScaler` from `sklearn.preprocessing` to standardize the features. This step ensures that individual features are adjusted to have a mean of zero and a standard deviation of one, which is particularly important for classifiers sensitive to feature magnitudes. The architecture for each classifier pipeline is structured as follows:

- **Logistic Regression Pipeline:**
$$\text{Pipeline} \rightarrow (\text{StandardScaler} \rightarrow \text{Logistic Regression})$$

- **Random Forest Pipeline:**
$$\text{Pipeline} \rightarrow (\text{StandardScaler} \rightarrow \text{Random Forest})$$

- **Gradient Boosting Pipeline (XGBoost):**
$$\text{Pipeline} \rightarrow (\text{StandardScaler} \rightarrow \text{XGBoost})$$

- **SVM Pipeline:**
$$\text{Pipeline} \rightarrow (\text{StandardScaler} \rightarrow \text{SVM})$$

3.2.4 Workflow Summary

The training workflow systematically proceeds as follows:

$$\text{Preprocessed Features} \rightarrow \text{Model Training} \rightarrow \text{Predictions}$$

This structured workflow leverages normalized and standardized feature sets, yielding predicted probabilities and class labels from each trained classifier. By conducting this comparative evaluation, we aim to identify the most effective methodologies for the classification task while illuminating the underlying dynamics of each classifier across the diverse multi-modal biomedical dataset.

4 Experiments

4.1 Experimental Settings

This section provides a comprehensive overview of the experimental settings utilized in this study, detailing the dataset specifications, evaluation metrics, baseline models, and implementation details to facilitate understanding and reproducibility.

4.1.1 Datasets and Preprocessing

The experiments utilized the AI-READI mini dataset, which comprises multi-modal biomedical data collected from 100 participants (one participant, ID 4034, was excluded per the task specification, leaving 99 participants). The dataset includes the following components:

- **Wearable Activity Monitor Data:** This includes raw JSON signals from various activity monitors capturing vital signals:
  - Heart rate
  - Oxygen saturation
  - Respiratory rate
  - Stress levels
  - Sleep patterns
  - Physical activity metrics
  - Active calorie expenditure
- **Continuous Glucose Monitor (CGM) Data:** Raw time series data processed using the Glucose360 library to extract clinical glycemic-variability features.
- **Clinical Data from OMOP CDM:** Clinical features are aggregated from several key tables, including:
  - person.csv
  - condition_occurrence.csv
  - measurement.csv
  - observation.csv
  - procedure_occurrence.csv
  - visit_occurrence.csv

To maintain analytical integrity, multiple preprocessing steps were executed, including the integration of disparate datasets into a cohesive feature framework. A critical part of preprocessing involved preventing label leakage by excluding diabetes-related indicators (diabetes/pre-diabetes diagnosis codes and HbA1c measurements) from the clinical data features. Model performance was evaluated using stratified 5-fold cross-validation (random seed 42) over the full cohort, rather than a single held-out split, to obtain a more stable estimate of performance on this small dataset.

Table 1 summarizes the dataset statistics, highlighting the participant count and diabetes-classification distribution.

Attribute	Count	Percentage
Total participants	99	100%
Diabetic (pre-diabetes, oral-medication, or insulin-dependent)	68	68.7%
Non-Diabetic (Healthy)	31	31.3%

Table 1: Dataset Statistics

4.1.2 Evaluation Metrics

Classifier performance was evaluated using several crucial metrics, defined as follows:

- **AUROC (Area Under the Receiver Operating Characteristic Curve):** Measures the model's ability to effectively distinguish between diabetic and non-diabetic participants.
- **Macro F1 Score:** Provides a harmonic mean of precision and recall, averaged across all classes, making it particularly useful for imbalanced datasets.
- **Accuracy:** Represents the proportion of correct predictions relative to the total number of predictions made by the model.
- **Confusion Matrix:** Visualizes the counts of true positives, true negatives, false positives, and false negatives, facilitating a breakdown of classification results for the overall test set and the insulin-dependent subgroup.

4.1.3 Baselines

To establish benchmark performance levels, the following baseline classifiers were implemented:

- **Logistic Regression**
- **Random Forest Classifier**
- **Gradient Boosting Classifier (XGBoost)**
- **Support Vector Machine (SVM)**

These baselines serve as reference points for evaluating the performance of the proposed classification methods and to contextualize the obtained results.

4.1.4 Implementation Details

The experimental models' implementation involved several critical components:

- Utilization of `StandardScaler` to standardize feature values, fit on each fold's training partition only.
- Stratified 5-fold cross-validation (`StratifiedKFold`, `shuffle=True`) to preserve the class distribution across folds.
- Default hyperparameters were retained for all baseline models to maintain consistency during performance evaluations.
- A fixed random seed of 42 was established across all stages of model training to support reproducibility of results.

4.2 Main Performance Comparison

To evaluate the classifiers' efficacy in distinguishing between diabetic and non-diabetic participants, we employed a multi-modal feature set derived from wearable activity monitors, Continuous Glucose Monitor (CGM) data processed via Glucose360, and clinical data organized according to the OMOP Common Data Model (CDM). The classifiers under consideration included Logistic Regression, Random Forest, Gradient Boosting (XGBoost), and Support Vector Machine (SVM).

Table 2 reports the mean of each metric across the 5 cross-validation folds.

Classifier	AUROC	Macro F1	Accuracy	Subgroup Accuracy
Logistic Regression	0.8262	0.6774	0.7263	1.0000
Random Forest	0.7732	0.6920	0.7574	1.0000
Gradient Boosting (XGBoost)	0.6852	0.5445	0.6253	1.0000
SVM	0.7914	0.6076	0.7263	1.0000

Table 2: Mean performance of baseline classifiers across 5-fold stratified cross-validation.

The comparative analysis demonstrates that Logistic Regression achieved the highest mean AUROC (0.8262), followed by SVM (0.7914) and Random Forest (0.7732). Random Forest achieved the highest mean accuracy (0.7574) and Macro F1 score (0.6920). Gradient Boosting (XGBoost) underperformed the other baselines across every metric on this small dataset, with a mean AUROC of 0.6852 and accuracy of 0.6253, suggesting it overfits more readily given the limited sample size (99 participants).

All four classifiers achieved a mean subgroup accuracy and Macro F1 of 1.0 for the insulin-dependent subgroup (n=6) across folds. Subgroup AUROC and a fold-level subgroup confusion matrix could not be reliably computed: with only 6 insulin-dependent participants split across 5 folds, individual test folds frequently contained too few subgroup examples (or examples of only one class) to compute these statistics, and we report this limitation rather than an unsupported number.

5 Conclusion

This study addresses the urgent need for effective diabetes classification given the increasing prevalence of the condition and the limitations of traditional methods. We proposed a novel methodology leveraging multimodal biomedical data, integrating wearable activity monitors, Continuous Glucose Monitoring (CGM) data, and clinical features to enhance prediction accuracy. The empirical results indicated that classifiers such as Logistic Regression and Random Forest achieved commendable performance metrics, with significant improvements over traditional approaches. Future research will aim to further refine data integration and feature extraction processes, explore additional data modalities, and address practical deployment challenges in real-time clinical settings, thereby contributing to personalized diabetes management and improving patient outcomes.

References

[1] F. Pedregosa et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011.

[2] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. KDD, 2016.

[3] L. Breiman. Random Forests. Machine Learning, 45(1):5–32, 2001.

[4] M. Abadi et al. TensorFlow: A System for Large-Scale Machine Learning. OSDI, 2016.

[5] A. Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS, 2019.

[6] G. Ke et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NeurIPS, 2017.

[7] A. Géron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, 2022.

[8] INRIA. Machine Learning in Python with scikit-learn (MOOC). fun-mooc.fr, 2022.

[9] Community resources. Hands-on machine learning tutorials. 2023.

[10] Community resources. Machine learning workshops and real-world datasets. 2023.

[11] Community resources. Interactive learning platforms for machine learning education. 2023.

[12] International Diabetes Federation. IDF Diabetes Atlas, 10th edition. 2021.

[13] S. Kaufman, S. Rosset, and C. Perlich. Leakage in Data Mining: Formulation, Detection, and Avoidance. KDD, 2011.

[14] Survey. Machine learning approaches for diabetes risk prediction. 2023.

[15] Survey. Wearable sensor data for diabetes monitoring and classification. 2023.