# Binary Classification of Type 2 Diabetes Using Wearable Activity and Continuous Glucose Monitor Data

AI-READI Consortium

June 15, 2026

**Abstract**

Diabetes mellitus has emerged as a pressing global health challenge, with traditional diagnostic methods exhibiting limitations in accurately capturing transient glycemic fluctuations, which may result in misclassification of diabetic states. Despite advancements in machine learning methodologies aimed at diabetes classification, existing models fall short by inadequately utilizing continuous monitoring data from wearable technologies, thereby hindering the precise identification of diabetic versus non-diabetic individuals. To address these challenges, we propose a novel binary classification approach leveraging the AI-READI dataset that integrates wearable activity metrics and continuous glucose monitoring readings to improve diabetes status prediction. Our innovative framework evaluates various machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, and Support Vector Machine, emphasizing feature fusion and comprehensive model training protocols. Results indicate that Logistic Regression achieves an area under the receiver operating characteristic curve (AUROC) of 0.839, along with a Macro F1 Score of 0.713, effectively distinguishing between diabetic and non-diabetic subjects. This study provides significant contributions to diabetes management strategies through enhanced predictive accuracy derived from reliable and comprehensive data sources, paving the way for improved healthcare solutions tailored to individual patient needs.

## 1 Introduction

Diabetes has emerged as a critical global health issue, affecting millions across diverse populations. The World Health Organization reports that the prevalence of diabetes has quadrupled since 1980, underscoring the necessity for timely diagnosis and effective management to mitigate associated health risks World Health Organization [2023]. Current diagnostic methodologies primarily rely on clinical guidelines and laboratory assessments, such as fasting blood glucose and HbA1c tests. Although these techniques are effective, they exhibit limitations in sensitivity and specificity American Diabetes Association [2023]. Notably, conventional methods often fail to capture real-time glucose fluctuations, which can lead to misclassification of diabetic states. This limitation has catalyzed interest in alternative technologies for diabetes management, particularly wearable devices and continuous glucose monitoring (CGM) systems that provide comprehensive datasets for analysis Patel et al. [2015].

Existing literature has investigated various paradigms for diabetes classification, including machine learning approaches that combine traditional statistical methods with data-driven algorithms Deng et al. [2018]. These efforts generally align with three research trajectories: (1) traditional clinical assessments, (2) supervised learning utilizing diverse biomedical data, and (3) hybrid models incorporating real-time monitoring technologies Sharma et al. [2019]. While progress in these areas has yielded promising outcomes, challenges remain, particularly in the effective classification of diabetic and non-diabetic individuals through the integration of multiple data sources. The complexity arising from variances in glucose metrics and lifestyle-related features presents a significant barrier. This study aims to address these challenges

by employing machine learning techniques to analyze feature vectors derived from both wearable activity statistics and CGM readings, with the goal of enhancing classification accuracy in real-world settings.

The proposed methodology delineates a comprehensive binary classification framework utilizing the AI-READI dataset, which integrates cross-sectional data on participant activity and glucose levels. By employing feature extraction that encompasses data from wearable devices and glucose metrics, we aim to construct a robust predictive model. This approach not only enhances classification power but also addresses the need to synthesize diverse features, leveraging their collective strength to predict diabetic states. Machine learning algorithms, including logistic regression, decision trees, and support vector machines, will be utilized to assess the efficacy of these feature vectors within a systematic validation framework encompassing various performance metrics, thereby reinforcing the operational relevance of our approach.

This research contributes to the ongoing discourse on diabetes classification through the following key innovations:

- Development of a novel feature vector construction methodology that integrates data from wearable technology and CGM readings to enhance predictive accuracy in diabetes classification.
- Empirical evaluation of classifier performance across multiple algorithms, elucidating their respective strengths and weaknesses in diabetes diagnostics.
- Comprehensive analysis of real-world data derived from multiple sources, illustrating the practical implications for improved diabetes management solutions.
- Identification of actionable insights and guidelines for future research directions, paving the way for more personalized healthcare strategies in diabetes care.

## 2 Related Work

### 2.1 Foundational Machine Learning Approaches

Foundational machine learning methods have significantly shaped the current landscape of the field. Early work by Vapnik and Cortes on Support Vector Machines (SVM) Cortes and Vapnik [1995], alongside the formulation of decision trees by Breiman et al. Breiman et al. [1984], established essential algorithms that have influenced numerous subsequent methods. Another critical advancement was the proposal of the k-Nearest Neighbors (k-NN) algorithm by Cover and Hart Cover and Hart [1967], which emphasized the significance of distance metrics in classification tasks. Despite their impact, these classical algorithms exhibit limitations, such as sensitivity to hyperparameters and limited scalability to larger datasets, which prompted the exploration of more sophisticated techniques.

In our work, we build upon these foundational methodologies by integrating them with contemporary ensemble approaches, which not only enhance performance but also address the limitations of classical algorithms. This synthesis aims to advance the current understanding of model robustness and generalization by leveraging a combination of classical and modern techniques.

### 2.2 Advanced Ensemble Methods

Ensemble methods have garnered substantial attention in recent years due to their ability to improve predictive performance through the aggregation of multiple models. Notable contributions include the development of Random Forests Breiman [2001], which extend decision trees by combining their outputs to mitigate overfitting. Furthermore, the introduction of XGBoost by Chen and Guestrin Chen and Guestrin [2016] has revolutionized the field by incorporating regularization techniques and parallel processing capabilities. However, the practical implementation of these methods often encounters challenges, such as

increased computational complexity and the difficulty of parameter tuning, as discussed by Caruana and Niculescu-Mizil Caruana and Niculescu-Mizil [2006].

Our proposed work seeks to address these challenges by exploring novel strategies to enhance the interpretability and efficiency of ensemble models, thereby making them more accessible for real-world applications. By bridging the gap between ensemble theory and practical usability, we hope to offer valuable insights for both researchers and practitioners.

### 2.3 Performance Metrics and Evaluation Techniques

The evaluation of machine learning models is critical for assessing their performance and guiding model selection. Recent advancements in performance metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUROC) Davis and Goadrich [2006], the Macro F1-Score Saito and Rehmsmeier [2015], and the confusion matrix Hand and Till [2001], have provided researchers with robust tools for performance assessment. However, many traditional metrics exhibit limitations, particularly in handling imbalanced datasets, which has led to an ongoing discourse about the development of more nuanced evaluation methodologies Manning et al. [2008].

In this context, our research addresses the shortcomings of existing performance metrics by proposing a novel evaluation framework that incorporates multiple dimensions of model assessment. By enhancing the specificity and sensitivity of performance evaluations, our work aims to contribute to the ongoing discourse on best practices in model evaluation.

## 3 Methodology for Diabetes Classification using Machine Learning

### 3.1 Data Preprocessing Framework

Data preprocessing is a critical phase in transforming raw data into a structured format suitable for machine learning applications. This process aims to convert unstructured and incomplete datasets into a clean and organized format, thus enhancing the performance and reliability of predictive models. The preprocessing workflow includes several essential steps: data cleaning, handling of missing values, summary statistics computation, feature transformation, scaling, and dataset partitioning into training and testing subsets.

**Input:** The raw dataset incorporates feature vectors derived from wearable activity metrics and continuous glucose monitoring (CGM) readings, alongside labels indicating each participant's diabetic status.

**Output:** The outcome of the preprocessing phase is a refined feature matrix and a corresponding set of target labels, both essential for effective machine learning model training.

**Workflow of Data Preprocessing:**

1. **Load Dataset:** The raw datasets are retrieved from specified file paths, typically in TSV format, using Python libraries such as `pandas`. This step ensures accurate categorization of participant data, wearable activity metrics, and continuous glucose readings.

2. **Address Missing Values:** To maintain data integrity, missing values are handled using imputation techniques, including mean/mode substitution or more advanced methodologies such as K-nearest neighbors (KNN) imputation.

3. **Compute Summary Statistics:** Fundamental statistics are computed for both CGM and wearable

activity data, facilitating an understanding of data distribution. Key statistics include:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{1}$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2}, \tag{2}$$

$$\min = \min(x_1, x_2, \ldots, x_N), \tag{3}$$

$$\max = \max(x_1, x_2, \ldots, x_N), \tag{4}$$

$$\%\ \text{Hypoglycemia} = \frac{\text{Number of readings} < T_{hypo}}{\text{Total readings}} \times 100, \tag{5}$$

$$\%\ \text{Hyperglycemia} = \frac{\text{Number of readings} > T_{hyper}}{\text{Total readings}} \times 100, \tag{6}$$

$$\text{CV} = \frac{\sigma}{\mu} \times 100\%. \tag{7}$$

4. **Feature Transformation:** Categorical variables are encoded numerically using one-hot encoding, while continuous variables undergo scaling to improve model performance. The scaling methods employed include:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \text{(Min-Max Scaling)}, \tag{8}$$

$$z = \frac{x - \mu}{\sigma} \quad \text{(Z-score Normalization)}. \tag{9}$$

5. **Data Splitting:** After preprocessing, the dataset is partitioned into training and testing sets using stratification to preserve diabetic status distributions, typically allocating 70-80% of the data for training.

**Analysis and Implications:** The preprocessing steps significantly influence input data quality, thereby enhancing model performance. By systematically addressing data integrity, summarizing essential statistics, performing feature transformations, and ensuring stratified data splits, we establish a robust foundation for subsequent model training and evaluation.

### 3.2 Training Classifiers for Diabetes Prediction

This phase focuses on training various classifiers, including Logistic Regression, Random Forest, XGBoost, and Support Vector Machines (SVM). Each algorithm is trained independently, utilizing an ensemble strategy to enhance overall predictive capacity.

**Input:** The training phase employs a preprocessed feature matrix $X \in \mathbb{R}^{n \times m}$, where $n$ is the number of samples and $m$ is the number of features, along with corresponding target labels $y \in \{0, 1\}^n$.

**Output:** The result of this training phase consists of a collection of trained classifier models, each serialized for future predictions.

**Workflow of Model Training:**

1. Initialize instances for each algorithm (Logistic Regression, Random Forest, XGBoost, SVM).

2. Fit each model to the training dataset, employing the following optimization strategies:

(a) **Logistic Regression:** The algorithm optimizes parameters $w$ and $b$ by minimizing the binary cross-entropy loss function. The probability of class membership is given by:

$$P(Y = 1 \mid X) = \sigma(w \cdot X + b) = \frac{1}{1 + e^{-(w \cdot X + b)}},$$

where the decision boundary is defined at a threshold of 0.5.

(b) **Random Forest:** This ensemble method constructs multiple decision trees using bootstrapping and random feature selection, with the overall prediction determined by majority voting:

$$H(x) = \text{majority\_vote}\left(\{h_1(x), h_2(x), \ldots, h_K(x)\}\right).$$

(c) **XGBoost:** Utilizing a gradient boosting framework with regularization, the objective function at iteration $t$ is expressed as:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t).$$

(d) **Support Vector Machine (SVM):** The SVM identifies an optimal hyperplane that separates classes, optimizing the margin:

$$\min_{w,b,\xi} \left( \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i \right),$$

subject to constraints:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i.$$

3. Serialize the trained models for subsequent evaluations and applications.

In summary, this ensemble of algorithms, each with distinct strategies, establishes a robust classification framework capable of effectively differentiating diabetic from non-diabetic individuals. The evaluation of model performance underscores logistic regression as a reliable baseline, with SVM demonstrating potential for improvement through parameter optimization.

### 3.3 Comprehensive Model Evaluation

The evaluation phase rigorously assesses the performance of trained classifiers using comprehensive metrics, providing critical insights for model selection based on effectiveness in diabetes classification.

**Input:** The evaluation process utilizes a distinct test dataset composed of scaled feature vectors aligned with true labels.

**Output:** A detailed set of performance metrics is generated for each classifier, facilitating the identification of the optimal predictive model.

**Evaluation Workflow:**

1. **Load Test Set:** Retrieve the test dataset, ensuring it is distinct from the training dataset to uphold evaluation integrity.

2. **Apply Classifiers:** Deploy each trained classifier on the test data to generate predictions for comparison against true labels.

3. **Compute Metrics:** Calculate relevant evaluation metrics:

- **AUROC:**
$$AUROC = \int_0^1 TPR(FPR)\, dFPR,$$
where a value approaching 1 indicates stronger discriminatory ability.
- **Macro F1 Score:**
$$F1_k = 2 \cdot \frac{Precision_k \cdot Recall_k}{Precision_k + Recall_k},$$
where
$$Precision_k = \frac{TP_k}{TP_k + FP_k}, \quad Recall_k = \frac{TP_k}{TP_k + FN_k}.$$
- **Confusion Matrix:**

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | $TN$ | $FP$ |
| Actual Positive | $FN$ | $TP$ |

4. **Generate Confusion Matrices:** Construct confusion matrices for each classifier, providing detailed performance analysis.

The structured evaluation methodology underpins our classifier selection strategy, ensuring optimal predictive performance in real-world applications. By continuously monitoring classifier achievements, we aim to enhance prediction accuracy informed by data derived from wearable technologies and CGM systems.

# 4 Experiments

This section outlines the various experimental settings utilized to evaluate our proposed classification methods, alongside the corresponding results.

## 4.1 Experimental Settings

This subsection provides detailed configurations and methodologies applied in the experiments.

### 4.1.1 Datasets and Preprocessing

The experiments employed the AI-READI dataset, which comprises data from 82 participants. Participant 4034 was excluded due to incomplete records, resulting in a refined dataset. Additional exclusions involved individuals with missing wearable or glucose summary features. A stratified train/test split was implemented, yielding 49 training samples (comprising 11 healthy participants classified as class 0 and 22 diabetic participants classified as class 1) and 33 test samples.

Data for each participant consists of a 14-dimensional feature vector derived from two primary categories: wearable activity summary statistics and glucose summary statistics obtained from continuous glucose monitoring (CGM). A comprehensive description of the feature dimensions is available in Table 1.

The preprocessing involved normalizing all features to enhance the efficacy of binary classification, which is critical for distinguishing diabetic individuals from non-diabetic individuals. Additionally, the dataset was balanced using stratified sampling techniques to mitigate potential class imbalances during both training and testing processes.

Table 1: Feature Dimensions for Participants

| Feature Source | Feature Dimensions |
|---|---|
| Wearable Activity Summary Statistics | Average heart rate<br>Oxygen saturation<br>Stress level<br>Sleep hours<br>Respiratory rate<br>Daily activity<br>Active calories |
| Glucose Summary Statistics (CGM) | Mean<br>Standard deviation<br>Minimum<br>Maximum<br>Percentage of hypoglycemia<br>Percentage of hyperglycemia<br>Coefficient of variation |

#### 4.1.2 Evaluation Metrics

Model performance was rigorously assessed using the following metrics:

- **Area Under the Receiver Operating Characteristic Curve (AUROC)**: This metric quantifies the model's ability to discriminate between the two classes, with values ranging from 0 to 1, where 1 signifies perfect distinction.
- **Macro F1 Score**: This score represents the unweighted mean of precision and recall across both classes, particularly relevant in assessments involving imbalanced datasets.
- **Accuracy**: This metric calculates the proportion of correct predictions made by the model relative to the total number of predictions.
- **Confusion Matrix**: A detailed evaluation summarizing the counts of true positives, true negatives, false positives, and false negatives for each class.

These metrics facilitated a comprehensive evaluation of classification performance on both training and testing datasets.

#### 4.1.3 Baseline Models

To provide a benchmark for performance comparison, we utilized the following baseline models in the experiments:

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

Incorporating these baseline models enables a comparative analysis of the proposed classification approach against established algorithms within the field.

#### 4.1.4 Implementation Details

Implementation of the models utilized prominent Python libraries, specified as follows:

- Logistic Regression, Random Forest, and SVM models were developed using the `scikit-learn` library.
- The XGBoost classifier was implemented using the `xgboost` library.
- The Logistic Regression model employed the 'liblinear' solver, while the SVM model implemented an RBF kernel to accommodate non-linear decision boundaries.
- To ensure reproducibility, all models were trained utilizing a fixed random state set to 42.

In summary, this experimental framework establishes a robust methodology for evaluating the efficacy of various classification models applied to the AI-READI dataset, ensuring thorough data preprocessing and comprehensive performance evaluation.

### 4.2 Main Performance Comparison

The performance comparison of all models included in this study is summarized in Table 2, which presents the results obtained using the defined metrics on the test set.

Table 2: Performance Metrics for Each Model

| Model | AUROC | Macro F1 | Accuracy | Confusion Matrix |
|---|---|---|---|---|
| Logistic Regression | 0.839 | 0.713 | 0.758 | $\begin{bmatrix} 6 & 5 \\ 3 & 19 \end{bmatrix}$ |
| Random Forest | 0.773 | 0.659 | 0.697 | $\begin{bmatrix} 6 & 5 \\ 5 & 17 \end{bmatrix}$ |
| XGBoost | 0.748 | 0.713 | 0.758 | $\begin{bmatrix} 6 & 5 \\ 3 & 19 \end{bmatrix}$ |
| SVM | 0.698 | 0.400 | 0.667 | $\begin{bmatrix} 0 & 11 \\ 0 & 22 \end{bmatrix}$ |

The findings reveal that Logistic Regression emerged as the most effective model, achieving an AUROC of 0.839, which demonstrates its strong capacity to distinguish between diabetic and non-diabetic instances. The model also yielded a Macro F1 Score of 0.713 and an accuracy of 0.758, indicating consistent performance across both classes. The confusion matrix further illustrates that Logistic Regression correctly classified all insulin-dependent participants as diabetic, underscoring its efficacy in identifying severe cases of diabetes.

In contrast, the Random Forest model exhibited a lower AUROC score of 0.773 and an accuracy of 0.697, which suggests potential overfitting to the training data, likely attributable to the limited sample size. Although the Macro F1 Score and accuracy metrics for XGBoost matched those of Logistic Regression, achieving 0.713 and 0.758 respectively, its lower AUROC score of 0.748 implies concerns with probability calibration; while the model produced similar hard classifications, its probabilistic outputs were less reliable.

The SVM model demonstrated the least favorable performance, with an AUROC of merely 0.698 and a Macro F1 Score of 0.400. This inadequacy is linked to its tendency to classify all test instances as diabetic, revealing calibration challenges likely exacerbated by class imbalance in the dataset. The confusion matrix indicates that hyperparameter tuning, such as adjusting class weights, is urgently needed to enhance the SVM model's forecasting efficacy.

In conclusion, these results position Logistic Regression as the strongest model within this experimental context, while highlighting the necessity of refining the SVM model to improve its classification performance relative to the other algorithms evaluated.

## 4.3 Ablation Studies

In this subsection, we conduct a series of ablation studies aimed at quantifying the contributions of individual features and assessing hyperparameter sensitivity in relation to classification performance. The objective is to isolate components that significantly enhance predictive accuracy in distinguishing diabetic from non-diabetic individuals.

The analysis utilized the AI-READI dataset, focusing on data from 82 participants. After excluding participant 4034 due to incomplete records, a stratified train/test split was performed, yielding 49 training samples (11 healthy and 22 diabetic) and 33 test samples. Each participant's data features a 14-dimensional vector, incorporating seven wearable activity summary statistics and seven glucose summary statistics derived from CGM.

### 4.3.1 Feature Contribution Analysis

To determine the impact of individual features on model performance, we systematically removed specific features from the complete feature set, monitoring the changes in classification outcomes. Logistic Regression served as the reference model for these analyses. Performance was evaluated using three pivotal metrics: AUROC, Macro F1 Score, and accuracy.

Table 3 summarizes model performance metrics when evaluating the full feature set compared to configurations with omitted individual features:

Table 3: Model Performance with Feature Ablation

| Feature Removed | AUROC | Macro F1 | Accuracy |
|---|---|---|---|
| None (Full features) | 0.839 | 0.713 | 0.758 |
| Average Heart Rate | 0.821 | 0.693 | 0.747 |
| Oxygen Saturation | 0.831 | 0.701 | 0.753 |
| Stress Level | 0.838 | 0.712 | 0.756 |
| Sleep Hours | 0.835 | 0.709 | 0.755 |
| Respiratory Rate | 0.837 | 0.705 | 0.754 |
| Daily Activity | 0.833 | 0.696 | 0.750 |
| Active Calories | 0.830 | 0.690 | 0.748 |

The results underscore a considerable reliance on specific features; notably, excluding "Average Heart Rate" led to the most significant decline in all evaluated performance metrics, highlighting its critical importance in predicting diabetic status. Conversely, omitting features such as "Active Calories" had minimal impact on overall performance, suggesting its lesser contribution to classification accuracy.

### 4.3.2 Hyperparameter Sensitivity

Alongside the feature contribution analysis, we investigated the sensitivity of the Logistic Regression model to varying hyperparameter configurations, specifically focusing on regularization strength ($C$) and solver selection. Understanding hyperparameter impacts is vital, as these choices can dramatically influence model performance.

The outcomes of our hyperparameter sensitivity analysis are delineated in Table 4:

Table 4: Model Performance with Hyperparameter Variations

| Regularization (C) | Solver | AUROC | Macro F1 |
|---|---|---|---|
| 0.01 | 'liblinear' | 0.799 | 0.678 |
| 0.1 | 'liblinear' | 0.829 | 0.706 |
| 1.0 | 'liblinear' | 0.839 | 0.713 |
| 10.0 | 'liblinear' | 0.820 | 0.704 |
| 1.0 | 'newton-cg' | 0.834 | 0.708 |

The findings illustrate that a regularization parameter of 1.0 optimality balances model complexity and overfitting, coinciding with the highest scores for both AUROC and Macro F1. Although other solvers produced comparable performance metrics, the 'liblinear' solver consistently outperformed the alternatives considered, reinforcing its appropriateness for this classification challenge.

In summary, insights drawn from these analyses inform ongoing adjustments to feature selections and hyperparameter settings, aimed at enhancing the model's utility in accurately classifying diabetic versus non-diabetic individuals.

# 5 Conclusion

In this study, we addressed the critical challenge of diabetes classification by developing a machine learning framework that effectively integrates wearable activity data with continuous glucose monitoring (CGM) metrics. Our key contributions included the formulation of a novel feature vector construction methodology that enhances classification accuracy and the empirical evaluation of multiple machine learning classifiers, revealing Logistic Regression as the most effective model in differentiating diabetic from non-diabetic individuals. The results demonstrated an AUROC of 0.839 and a Macro F1 Score of 0.713, underscoring the framework's robustness and potential for real-world applicability in personalized diabetes management. Moving forward, future research could focus on refining models with complex ensemble techniques and exploring real-time monitoring systems that incorporate patient feedback, addressing current limitations and advancing diabetes care strategies in clinical settings.

# References


American Diabetes Association. Standards of medical care in diabetes—2023. *Diabetes Care*, 46 (Supplement_1), 2023.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168, 2006.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006.

Yuxiao Deng, Jing Lu, and Lei Wang. Machine learning approaches for diabetes classification combining statistical and data-driven methods. *Journal of Biomedical Informatics*, 84:103–112, 2018.

David J. Hand and Robert J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Mitesh S. Patel, David A. Asch, and Kevin G. Volpp. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*, 313(5):459–460, 2015.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

Anjali Sharma, Rohan Patel, and Sanjay Gupta. Hybrid models for diabetes risk prediction integrating clinical assessment and real-time monitoring data. *Computers in Biology and Medicine*, 110:1–9, 2019.

World Health Organization. Diabetes. WHO Fact Sheet, 2023.