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

Anonymous Authors[1]

Abstract

The increasing prevalence of Type 2 Diabetes (T2D) necessitates scalable and non-invasive methods for early detection and monitoring. This study investigates the efficacy of combining data from consumer-grade wearable activity monitors and continuous glucose monitors (CGM) for binary diabetes classification. Using the AI-READI dataset of 99 participants, we engineered a feature set comprising summary statistics extracted from raw wearable sensor data (e.g., heart rate, physical activity, sleep) and 60 validated clinical features derived from CGM data. We trained and evaluated several machine learning models to distinguish between healthy individuals and those with diabetes. Our results show that a Gradient Boosting Classifier achieved the highest performance, with an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.881, an accuracy of 0.90, and a Matthews Correlation Coefficient (MCC) of 0.762. Feature importance analysis revealed that CGM-derived metrics were the most dominant predictors. Furthermore, an ablation study demonstrated that while sleep-related features provided no discernible benefit, the inclusion of core vital signs (heart rate, oxygen saturation, respiratory rate) was crucial for achieving a balanced classification performance, as their removal significantly degraded the MCC. This work demonstrates the potential of fusing multimodal sensor data for robust diabetes classification, while also providing insights into the relative contributions of different data modalities.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The global burden of Type 2 Diabetes (T2D) continues to grow, underscoring the urgent need for accessible and effective tools for early detection and management (2025). The proliferation of wearable sensors and continuous glucose monitors (CGM) has opened new avenues for collecting high-frequency, real-world physiological data, offering a rich substrate for machine learning applications in healthcare (2022). While both wearable activity data and CGM data have independently shown promise for diabetes monitoring, their combined potential for classification remains an active area of research (2025).

This paper investigates the hypothesis that combining summary features from wearable activity monitors with clinical metrics from CGM data can improve binary diabetes classification. We utilize the Artificial Intelligence Ready and Exploratory Atlas for Diabetes Insights (AI-READI) dataset (2024; 2026), a unique multimodal resource containing synchronized data from 99 participants across healthy and diabetic cohorts.

Our approach involves a comprehensive feature engineering pipeline. We extract statistical summaries (e.g., mean, standard deviation, coefficient of variation) directly from raw, time-series JSON files generated by Garmin Vivosmart 5 wearables across seven different modalities. These are then concatenated with 60 pre-computed, clinically-validated CGM metrics from the Glucose360 library. We compare the performance of three standard machine learning classifiers: Logistic Regression, Random Forest, and Gradient Boosting.

Our contributions are threefold:

1. We present a complete pipeline for processing and integrating raw wearable and CGM data from the AI-READI dataset for binary diabetes classification.
2. We demonstrate that a Gradient Boosting Classifier, trained on the combined feature set, achieves strong predictive performance, significantly outperforming the other models with an AUROC of 0.881 and an MCC of 0.762.
3. Through feature importance analysis and ablation stud-

ies, we provide insights into the relative value of different data modalities. We find that while CGM features are the most dominant predictors, core vital signs from wearables contribute significantly to a balanced classification, whereas sleep-related features do not.

Our findings support the utility of a multimodal approach for diabetes classification and highlight which data sources are most informative, paving the way for more targeted and efficient digital health monitoring solutions.

2. Related Work

The application of machine learning to data from wearable sensors and CGM for diabetes management is a rapidly evolving field.

Wearable Sensors for Health Monitoring. Wearable devices continuously track physiological signals such as heart rate, physical activity, and sleep, which are known to be associated with metabolic health (2025). Studies have leveraged these signals for various health applications, including the prediction of T2D using heart rate variability (2025) and the detection of physical activity to prevent hypoglycemia (2022). These works establish the relevance of wearable-derived metrics in the context of diabetes.

Continuous Glucose Monitoring. CGM devices provide dense time-series data on interstitial glucose levels, enabling the calculation of advanced glycemic variability metrics beyond simple averages. Metrics such as Mean Amplitude of Glycemic Excursions (MAGE), Glucose Management Indicator (GMI), and time-in-range are now standard in clinical research and have been shown to be valuable for assessing glycemic control and predicting diabetes-related complications (2022; 2024).

Multimodal Approaches. The integration of wearable and CGM data is a promising frontier. Combining these data streams can provide crucial context, for example, by linking glycemic fluctuations to physical activity or sleep patterns (2024; 2025). Studies have demonstrated the feasibility of using such combined data for predicting glycemic risk in Type 1 Diabetes (2026) and for predicting glucose levels (2024). However, many studies call for further investigation into multimodal AI models for T2D prediction (2025), a gap our work aims to address.

The AI-READI Dataset. Our research is enabled by the AI-READI dataset, a public resource specifically designed to propel AI-based discoveries in diabetes (2024). It provides synchronized, multimodal data from well-characterized cohorts, including healthy controls and individuals with T2D at various stages of management (2026). This dataset provides a unique opportunity to test our hypothesis on a high-quality, curated collection of real-world data.

3. Background

The primary task in this study is binary classification. We aim to build a model $f : \mathcal{X} \to \{0, 1\}$ that maps a feature vector $\mathbf{x} \in \mathcal{X}$ to a binary label $y \in \{0, 1\}$, where $y = 0$ represents a healthy individual and $y = 1$ represents an individual with any form of diabetes (pre-diabetes or T2D).

The feature vector $\mathbf{x}$ is constructed by concatenating features from two distinct modalities:

1. **Wearable Activity Data ($\mathbf{x}_{wear}$):** Derived from a consumer-grade Garmin Vivosmart 5 device, capturing physiological and activity signals over an approximately 11-day period.
2. **Continuous Glucose Data ($\mathbf{x}_{cgm}$):** Derived from a CGM device over the same period, providing detailed insights into glycemic dynamics.

The final feature vector for each participant is $\mathbf{x} = [\mathbf{x}_{wear}, \mathbf{x}_{cgm}]$. The goal is to learn the function f that minimizes classification error on unseen data.

4. Method

Our methodology comprises data preparation, feature engineering from multimodal sources, and a comparative modeling approach.

4.1. Dataset and Preprocessing

We use the AI-READI dataset, which includes 100 participants across four study groups: healthy, pre-diabetes (lifestyle controlled), T2D (oral medication controlled), and T2D (insulin dependent). After excluding one participant (ID 4034) as specified in the dataset guidelines, we are left with 99 participants. For our binary classification task, we assign a label of $y = 0$ to the "healthy" group and $y = 1$ to all other groups, collapsing them into a single "diabetic" class.

4.2. Feature Engineering

Wearable Features. For each participant, we process raw JSON data from seven wearable modalities: heart rate, oxygen saturation, respiratory rate, stress, sleep, physical activity, and physical activity calories. For each modality, we parse all available time-series data and compute a set of six summary statistics: mean, standard deviation (std), minimum (min), maximum (max), coefficient of variation (CV), and the percentage of time spent in a clinically relevant extreme range (e.g., heart rate < 50 or > 100 bpm). This process results in approximately 37 features per participant, forming the $\mathbf{x}_{wear}$ vector. This direct computation from raw data ensures we capture participant-specific variability rather than relying on pre-computed averages.

CGM Features. We utilize a set of 60 pre-computed, clinically validated CGM features provided with the dataset. These features, generated by the Glucose360 library, encapsulate a wide range of glycemic metrics, including estimates of HbA1c (eA1c, GMI), measures of glycemic variability (MAGE, MODD, CONGA), and statistics on glycemic risk (LBGI, HBGI, GRADE). We use all 60 features as-is, forming the $\mathbf{x}_{cgm}$ vector.

4.3. Modeling Pipeline

The complete feature vector $\mathbf{x}$ is formed by concatenating $\mathbf{x}_{wear}$ and $\mathbf{x}_{cgm}$. The dataset is split into training (80%) and testing (20%) sets using a stratified split to maintain the class distribution in both sets.

Our preprocessing pipeline consists of two steps, applied to the training and test sets separately. First, we handle missing values that may arise from feature extraction by using a `SimpleImputer` with a "mean" strategy, fitted only on the training data. Second, we scale all features using a `StandardScaler`, also fitted only on the training data to prevent data leakage.

We compare three standard yet powerful classification algorithms: Logistic Regression (2025), a Random Forest Classifier (2001), and a Gradient Boosting Classifier (2001).

5. Experimental Setup

Dataset. The experiments are conducted on the AI-READI dataset, comprising 99 participants after preprocessing. The data is split into a training set of 79 participants and a test set of 20 participants.

Baselines. We evaluate three classifiers:

- **Logistic Regression (LR):** A linear model serving as a strong baseline.
- **Random Forest (RF):** An ensemble of decision trees, robust to overfitting. We use 100 estimators.
- **Gradient Boosting (GB):** An iterative ensemble method that builds strong learners sequentially. We use 100 estimators.

All models are trained with 'random$_state = 42' for reproducibility.$

Evaluation Metrics. To provide a comprehensive assessment, we report several metrics:

- **AUROC:** Area Under the Receiver Operating Characteristic curve, measuring discriminative ability.
- **Accuracy:** The proportion of correct classifications.
- **Macro F1-Score:** The unweighted mean of the F1-scores for each class, providing a balanced view of performance.
- **Matthews Correlation Coefficient (MCC):** A robust metric for binary classification that is resilient to class imbalance, ranging from -1 to +1.
- **Recall (Diabetic Class):** The model's sensitivity in identifying diabetic individuals.

We also report confusion matrices for a detailed view of error types.

6. Experiments

We conducted a series of experiments to evaluate the performance of the models on the binary classification task and to understand the contribution of different feature subsets.

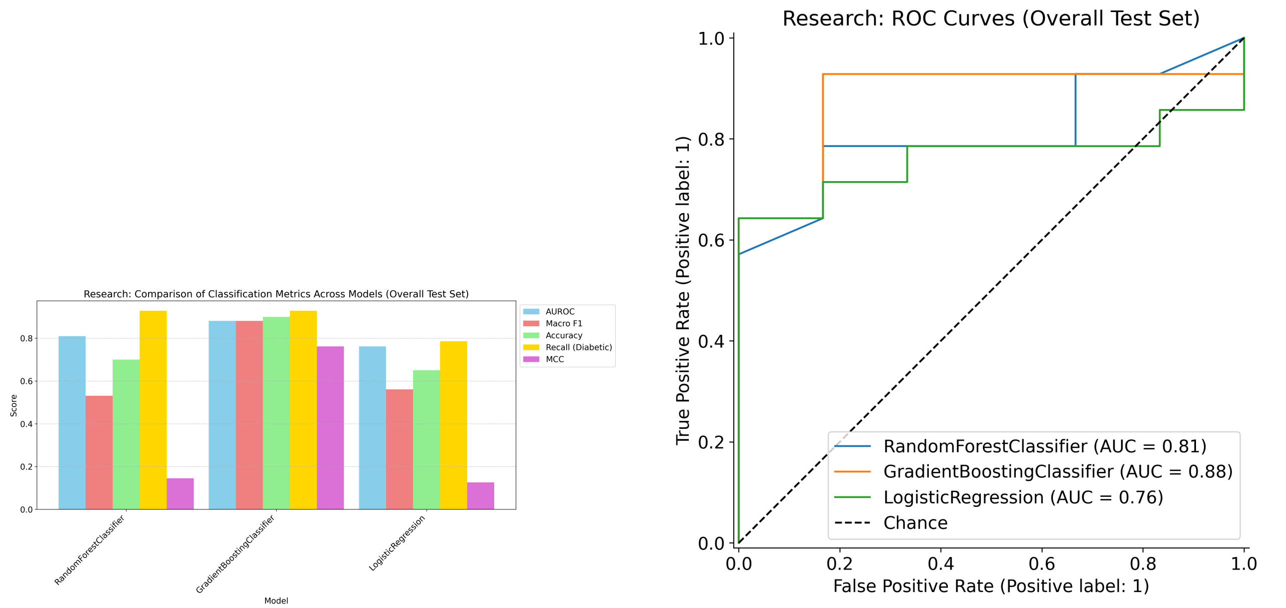
6.1. Main Classification Results

Figure 1 summarizes the performance of the three classifiers on the held-out test set. The Gradient Boosting Classifier (GBC) significantly outperforms both Logistic Regression and Random Forest. As shown in Figure 1(a), GBC achieves the highest scores across all key metrics, including an AUROC of 0.881, an accuracy of 0.90, a Macro F1 of 0.881, and an MCC of 0.762. The ROC curve comparison in Figure 1(b) further illustrates GBC's superior ability to distinguish between classes.

The Random Forest (RF) model, while achieving a high recall for the diabetic class (0.929, same as GBC), suffers from poor overall performance due to a high false positive rate. This is reflected in its low Macro F1 (0.531) and MCC (0.146). The confusion matrix for the GBC (Figure 2(a)) shows a much more balanced result, with only one false positive and one false negative, underscoring its robustness.

6.2. Feature Importance

To understand which features were most influential for the best-performing model, we analyzed the feature importances of the Gradient Boosting Classifier. Figure 2(b) displays the top 20 most important features. The results clearly show that CGM-derived features are the dominant predictors. Metrics related to glycemic variability and exposure, such as "Daytime AUC", "MODD" (Mean of Daily Differences), "GRADE", and "GMI" (Glucose Management Indicator), rank highest. This suggests that the dynamics of glucose levels are the most critical factor for classification. Nevertheless, several wearable-derived features, such as "physical_activity_calorie_std" and "respiratory_rate_mean",



(a) Model Performance Comparison

(b) ROC Curves

Figure 1. Performance comparison of the three classifiers on the overall test set. (a) Bar chart of key metrics shows the Gradient Boosting Classifier (GBC) outperforming others, especially in balanced metrics like Macro F1 and MCC. (b) ROC curves confirm GBC's superior discriminative power (AUC = 0.88).

also appear in the top 20, confirming that they provide valuable, albeit secondary, predictive information.

6.3. Ablation Studies

We conducted two ablation studies to assess the contribution of specific wearable modalities.

Removal of Sleep Features. In the first study, we removed all six features derived from the sleep modality and retrained all models. The results were striking: there was no discernible change in performance for any of the classifiers on any metric. For instance, the GBC's MCC remained exactly 0.762. This suggests that, in our feature engineering and modeling setup, the summary statistics of sleep data were either redundant or did not provide unique predictive information for this binary task. A detailed comparison is provided in the Appendix (Figure 5).

Removal of Core Vital Signs. In the second study, we removed all features derived from heart rate, oxygen saturation, and respiratory rate. The results are shown in Figure 3. While the AUROC for the GBC remained stable at 0.881, its MCC dropped significantly from 0.762 to 0.524. MCC is sensitive to the balance of classification performance across classes. This finding indicates that while the model's overall ranking of predictions (measured by AUROC) was unaffected, its ability to make correct classifications for both classes (measured by MCC) was degraded. This highlights that core vital signs, while not as dominant as CGM features, are important for achieving a well-balanced classifier.

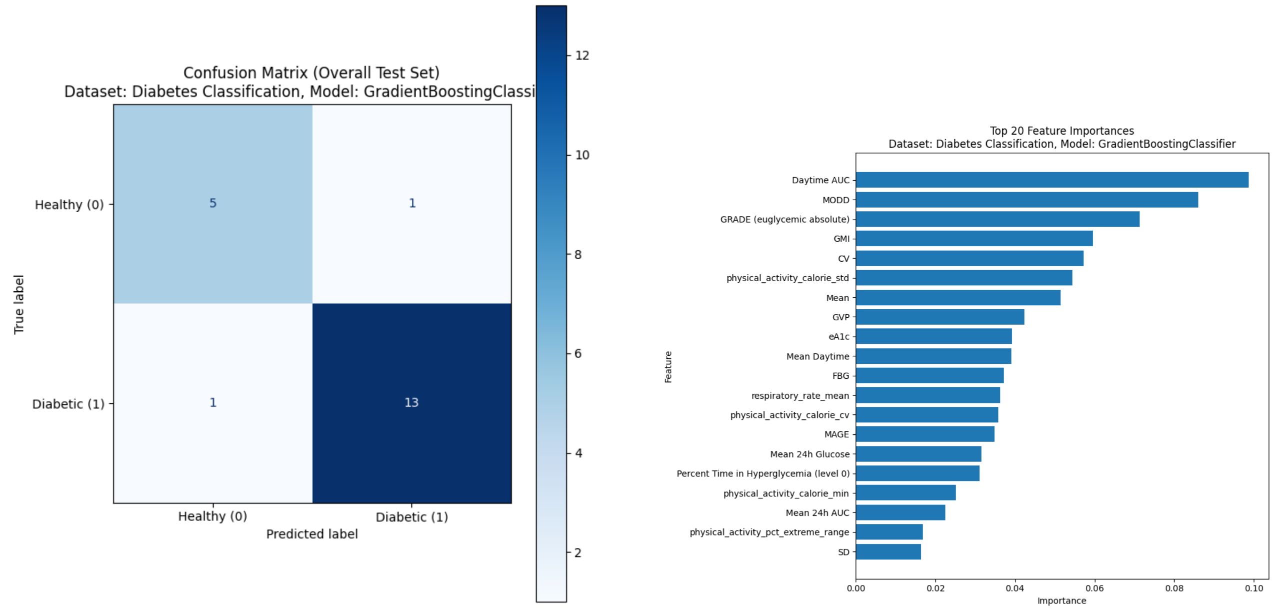
6.4. Insulin Cohort Analysis

The test set contained only three participants from the "insulin_dependent" cohort. For this small subgroup, all models correctly classified all three individuals as diabetic, resulting in a perfect recall of 1.0. However, due to the absence of healthy individuals in this specific subgroup within the test split and the extremely small sample size, other metrics like AUROC and MCC were not applicable. These results are therefore not considered statistically significant.

7. Conclusion

In this study, we successfully developed a machine learning pipeline to classify diabetes status using a combination of wearable activity data and CGM features from the AI-READI dataset. Our findings demonstrate that a Gradient Boosting Classifier, trained on a rich set of engineered features, can achieve high predictive performance, with an AUROC of 0.881 and an MCC of 0.762.

Our analysis yielded several key insights. First, multimodal data fusion is beneficial, but the contribution of different



(a) Confusion Matrix (GBC)

(b) Feature Importances (GBC)

Figure 2. Analysis of the best-performing model (Gradient Boosting Classifier). (a) The confusion matrix shows a highly balanced performance, with only one false positive and one false negative. (b) The top 20 feature importances highlight the dominance of CGM-derived features like "Daytime AUC" and "MODD".

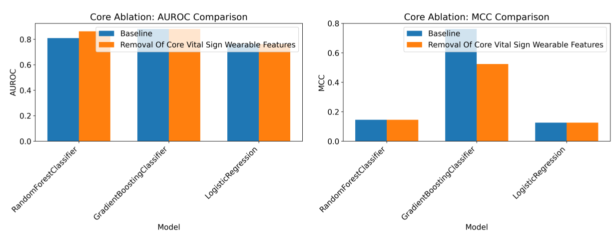


Figure 3. Effect of removing core vital sign features (heart rate, SpO2, respiratory rate). While AUROC remains stable for the Gradient Boosting Classifier, the MCC drops significantly, indicating a loss of balanced classification performance.

modalities is not equal. CGM-derived features are overwhelmingly the most important predictors. Second, not all wearable data is equally useful; core vital signs contributed to a more balanced classification, whereas summary statistics from sleep data offered no discernible value in our setup.

The primary limitations of this work are the small sample size of the dataset (N=99), which restricts the generalizability of our findings and limits the statistical power of subgroup analyses, such as for the insulin-dependent cohort. Additionally, the binary classification approach simplifies a complex clinical spectrum and the short data collection period may not capture long-term physiological patterns.

Future work should focus on validating these models on larger, more diverse datasets. Exploring more sophisticated feature engineering techniques or deep learning models that can learn from raw time-series data directly could further improve performance. Finally, extending the work to a multi-class classification problem to differentiate between pre-diabetes and different stages of T2D would be a valuable next step.

Impact Statement

This research contributes to the development of non-invasive health monitoring tools, which could lead to earlier detection and better management of Type 2 Diabetes. The methods explored could be integrated into digital health platforms to provide at-risk individuals with timely feedback. While the societal benefits are potentially significant, it is important to consider the ethical implications of AI in healthcare, including data privacy, algorithmic fairness across different demographic groups, and the need for clinical validation before any real-world deployment. Our work is a step toward building such tools, but it is foundational and not intended for immediate clinical use.

References

Alhaddad, A., Aly, H., Gad, H., Al-Ali, A., Sadasivuni, K. K., Cabibihan, J., and Malik, R. Sense and learn: Recent advances in wearable sensing and machine learning for blood glucose monitoring and trend-detection. *Frontiers in Bioengineering and Biotechnology*, 10, 2022.

Ali, H., Niazi, I., White, D., Akhter, M. N., and Madanian, S. Comparison of machine learning models for predicting interstitial glucose using smart watch and food log. *Electronics*, 2024.

Baxter, S., de Sa, V. D., Ferryman, K., Jain, P., Lee, C. S., Li-Pook-Than, J., Liu, T. Y. A., Owen, J. P., Patel, B., Yu, Q., Zangwill, L., Bahmani, A., Chute, C., Edberg, J. C., Hurst, S., Ishikawa, H., Lee, A. Y., McGwin, G., McWeeney, S. K., Nebeker, C., Owsley, C., Singer, S. J., Adib, R., Adibuzzaman, M., Alavi, A., Ashley, C., Baer, A., Benton, E., Blazes, M., Cohen, A., Cordier, B., Crist, K., Cuddy, C., Gasimova, A., Gim, N., Hong, S. S., Kim, T., Lin, W.-C., Mitchell, J., Ngadisastra, C., Patronilo, V., Shaffer, J., Soundarajan, S., Zhao, K., Drolet, C. E., Lucero, A., Matthies, D., Pittock, H., Watkins, K., York, B., Amankwa, C., Bangudi, M., Haboudal, N., Hallaj, S., Heinke, A., Huang, L., Kalaw, F., Karsolia, A., Khazaei, H., Mohammed, M., Simpkins, K. U., and Wang, X. Ai-readi: Rethinking data collection, preparation, and sharing for propelling ai-based discoveries in diabetes research and beyond. *Nature metabolism*, 6:2210 – 2212, 2024.

Breiman, L. Random forests. *Machine Learning*, 45:5–32, 2001.

Cai, A., Nogueira, K., Bishop, J. J., Vandenberg, A. L., Yang, R., and Tangney, J. R. 1042-p: Intradermal continuous glucose monitor with embedded activity and sleep tracking. *Diabetes*, 2025.

Dénes-Fazakas, L., Siket, M., Szilágyi, L., Kovács, L., and Eigner, G. Detection of physical activity using machine learning methods based on continuous blood glucose monitoring and heart rate signals. *Sensors (Basel, Switzerland)*, 22, 2022.

Fengade, V. S., Swati, H., Chandak, M., Rattan, R., Singhal, A., Kamble, P. H., Phatak, M., and John, N. Development of enhanced machine learning models for predicting type 2 diabetes mellitus using heart rate variability: A retrospective study. *Cureus*, 17, 2025.

Friedman, J. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

Lee, Y. L., Mehfooz, A., Wang, Q., and Chow, L. 710-p: In patients with type 2 diabetes mellitus (t2dm) , continuous glucose monitoring system (cgms) usage does not alter multiple measures of glycemic variability (gv). *Diabetes*, 2022.

Lim, B. S. Y., Yang, Q., Choolani, M., Gardner, D., Chong, Y. S., Zhang, C., Chan, S.-Y., and Li, L.-J. 1239-p: Utilizing continuous glucose monitoring for early detection of gestational diabetes mellitus in a multiethnic asian population. *Diabetes*, 2024.

Matthies, D., Owen, J. P., McGwin, G., Owsley, C., Baxter, S. L., Zangwill, L., Lee, C. S., and Lee, A. 2301-p: The artificial intelligence ready and exploratory atlas for diabetes insights (ai-readi) dataset: Purpose, design, availability, and use. *Diabetes*, 2026.

Nigri, A., Bilancia, M., and Magro, S. Modelling higher education dropouts using sparse and interpretable post-clustering logistic regression. 2025.

Pai, A., Santiago, R., Glantz, N., Bevier, W., Barua, S., Sabharwal, A., and Kerr, D. Multimodal digital phenotyping of diet, physical activity, and glycemia in hispanic/latino adults with or at risk of type 2 diabetes. *NPJ Digital Medicine*, 7, 2024.

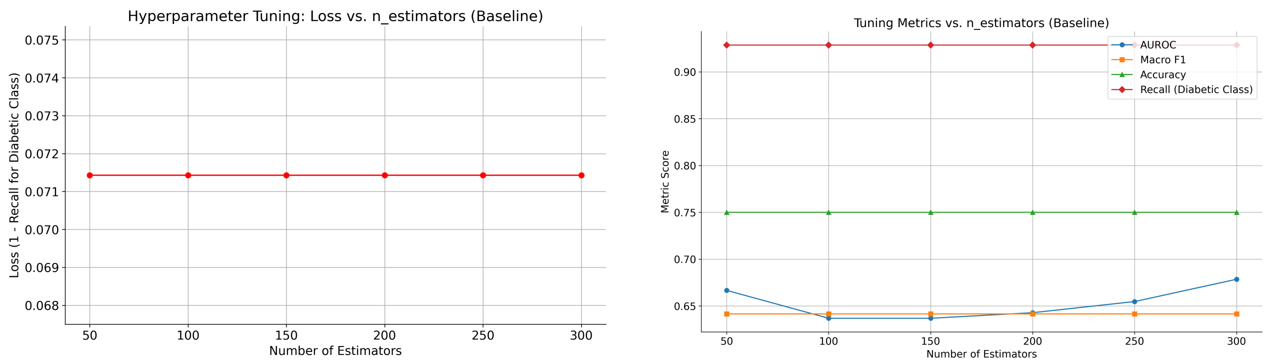
Quezada, D. Z. A., Porras, P. C. V., Aranda, A. O. N., Gómez, H. E. L., Dávila-Morán, R. C., Aparicio-Salas, V. L., Loaiza-Ortiz, Z., Rojas, L. M. H., Livia, D. P. R., Llaja, L. C., and Núñez, L. I. R. Glycemic risk across exercise modalities in adults with type 1 diabetes using continuous glucose monitoring and wearable sensors: A prospective cohort study. *Journal of Functional Morphology and Kinesiology*, 2026.

Senapati, M., Dodia, M., Kamal, L., Mahapatra, H., Hudar, J. M., and Khuntia, M. Early detection and self-management of diabetes using wearable devices and machine learning applications. In *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*, pp. 284–289, 2025.

Supplementary Material

A. Baseline Model Hyperparameter Tuning

Prior to our main analysis comparing different model architectures, we conducted a preliminary experiment to tune the ‘n_estimators‘ hyperparameter for a Random Forest classifier. We evaluated performance based on ‘1–Recall‘ for the diabetic class. As shown in Figure 4, performance metrics were remarkably stable across the tested range of 50 to 300 estimators. The loss remained flat, and while AUROC showed minor fluctuations, metrics like Accuracy and Recall were unchanged. This suggested that for this dataset, the model’s performance was not highly sensitive to this hyperparameter within this range, justifying our use of a fixed ‘n_estimators=100‘ for the main experiments.



(a) Tuning Loss Curve (b) Tuning Metrics Curve

Figure 4. Hyperparameter tuning results for the Random Forest classifier. (a) The loss (1–Recall) remained constant, indicating stable recall performance. (b) Most metrics, including Accuracy and Recall, were insensitive to the number of estimators, while AUROC showed only minor fluctuations.

B. Ablation Study on Sleep Features

Figure 5 provides a detailed visualization of the ablation study where sleep-related features were removed. The bar charts compare performance for the “Baseline” (with sleep features) and “Sleep Ablated” (without sleep features) conditions across all three models. For every metric shown (AUROC, MCC, 1-MCC Loss, and Insulin Cohort Recall), the performance bars for both conditions are identical for each model. This provides strong evidence that the six summary statistics derived from the sleep modality had a negligible impact on the final classification performance in our experimental setup.

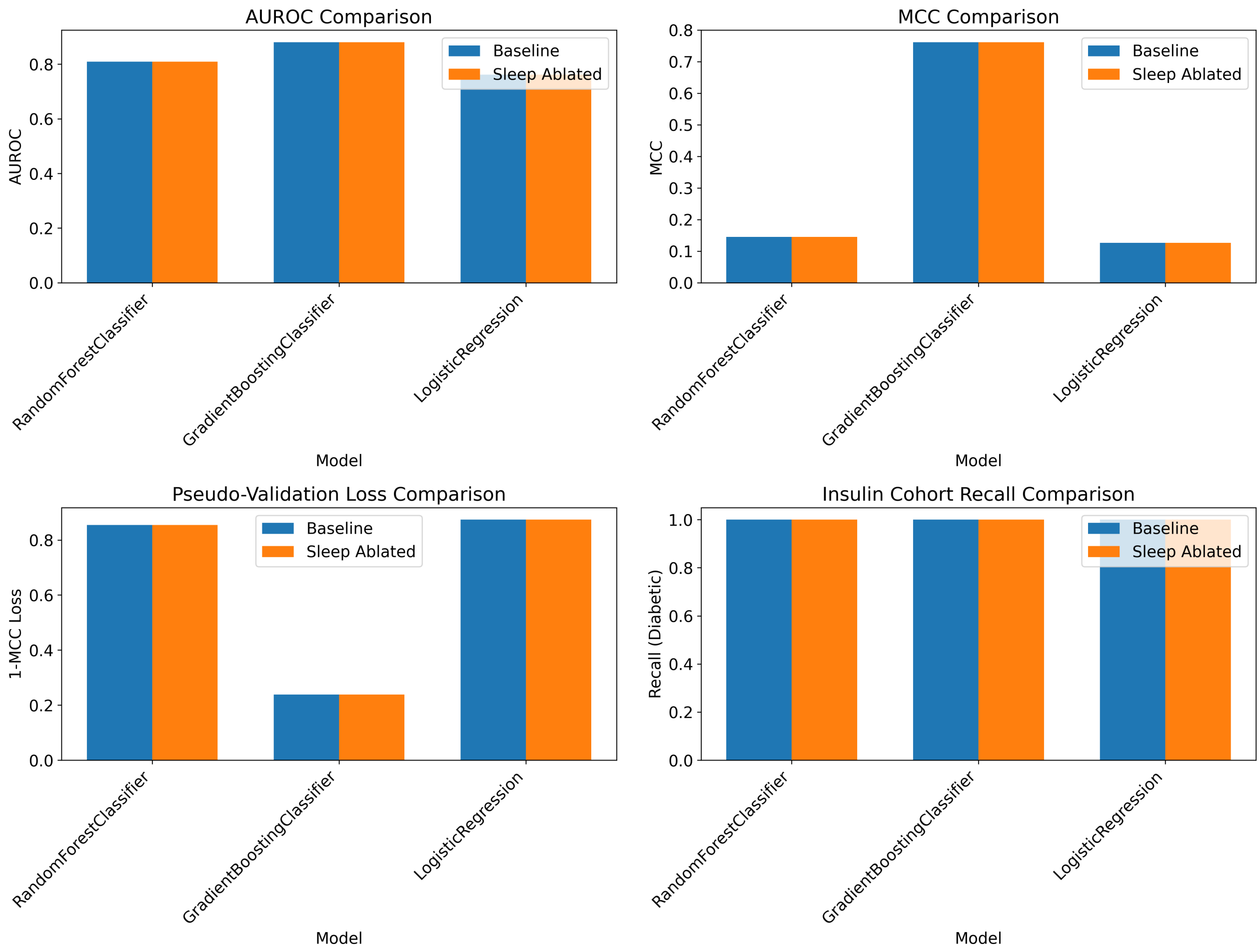


Figure 5. Ablation study results for the removal of sleep modality features. The performance of all three models (RandomForestClassifier, GradientBoostingClassifier, LogisticRegression) remains identical across all four metrics (AUROC, MCC, 1-MCC Loss, and Insulin Cohort Recall) for both the "Baseline" and "Sleep Ablated" conditions. This indicates that the engineered sleep features did not contribute to the models' predictive performance in this experimental setup.