Machine Learning Classification of Insulin Resistance and Beta-Cell Dysfunction Using Aggregated Continuous Glucose Monitoring Features

Dr Mazen M Salama
Senior Clinical Sentient - Dataemia - USA

Predicting insulin resistance and -cell dysfunction from continuous glucose monitoring data is critical for enabling early metabolic interventions, yet remains challenging due to the high dimensionality and non-linear dynamics of glycemic trajectories. To address this, we engineered 26 comprehensive features capturing basic statistics, variability indices, time-in-range metrics, trajectory parameters, distributional properties, and adherence signals from raw glucose readings. These features were evaluated using four machine learning classifiers: logistic regression, random forest, gradient boosting, and multilayer perceptron. Models were trained and validated on a clinical cohort of 457 subjects, with 44 fully labeled records partitioned into downstream training and validation sets. Our results demonstrate that predictive performance is task-dependent: gradient boosting achieved the highest accuracy for insulin resistance classification (0.772) with a macro-averaged F1-score of 0.740, while logistic regression proved most effective for -cell dysfunction prediction (accuracy: 0.711, ROC-AUC: 0.800). These findings indicate that carefully engineered time-series glucose features, paired with appropriately selected machine learning architectures, offer a robust and clinically actionable framework for the early detection of metabolic disorders.

I. INTRODUCTION

Continuous glucose monitoring (CGM) has fundamentally transformed diabetes management by providing high-frequency, real-time glycemic data. Beyond routine clinical care, CGM represents a rich physiological resource that can be leveraged to detect early metabolic dysregulation. Insulin resistance and β-cell dysfunction are the foundational pathophysiological mechanisms driving the progression to type 2 diabetes and prediabetes. Identifying these conditions early through non-invasive, data-driven approaches could enable timely lifestyle or pharmacological interventions, potentially halting disease progression before irreversible complications arise. Despite the abundance of CGM data, translating raw glucose trajectories into actionable metabolic insights remains a significant challenge. Glycemic patterns are inherently high-dimensional, noisy, and governed by complex non-linear dynamics influenced by diet, physical activity, circadian rhythms, and individual metabolic variability. Traditional diagnostic methods rely on static clinical tests or sparse fingerstick measurements, which fail to capture the dynamic interplay between insulin secretion and glucose clearance. Consequently, extracting robust, physiologically meaningful signals from continuous time-series data requires sophisticated feature engineering and adaptive modeling strategies that can handle the inherent complexity of glycemic trajectories without succumbing to overfitting or information loss.

To address this gap, we propose a machine learning framework that systematically transforms raw CGM readings into a compact yet comprehensive set of 26 engineered features. These features are explicitly designed to capture the multi-faceted nature of glycemic regulation, including basic statistical summaries, variability indices, time-in-range metrics, trajectory parameters, distributional properties, and adherence signals. By projecting high-dimensional glucose readings into this structured feature space, we reduce measurement noise while preserving clinically relevant physiological patterns. We then evaluate four distinct machine learning classifiers—logistic regression, random forest, gradient boosting, and a multilayer perceptron—to determine how different algorithmic inductive biases interact with these engineered representations. This multi-model comparison allows us to identify which architectures best capture the linear and non-linear dependencies inherent in metabolic dysregulation, ensuring that model selection is driven by task-specific performance rather than architectural preference.

We validate our framework on a clinical cohort comprising 457 subjects, with 44 fully labeled records partitioned into downstream training and validation sets to ensure rigorous assessment. By framing the prediction of insulin resistance and β-cell dysfunction as distinct classification tasks, we demonstrate that predictive performance is highly task-dependent. Gradient boosting emerges as the optimal architecture for classifying insulin resistance, achieving an accuracy of 0.772 and a macro-averaged F1-score of 0.740, while logistic regression proves most effective for β-cell dysfunction prediction, yielding an accuracy of 0.711 and a ROC-AUC of 0.800. These findings underscore that carefully engineered time-series glucose features, when paired with appropriately selected machine learning architectures, provide a robust and clinically actionable framework for the early detection of metabolic disorders. Our results establish a scalable pathway for leveraging routine CGM data to identify at-risk individuals before clinical diagnosis, ultimately supporting proactive metabolic healthcare.

II. METHODS

A. Data acquisition and preprocessing

The study utilized a clinical cohort of 457 subjects who underwent continuous glucose monitoring (CGM) over a standardized observation period. Raw CGM readings, sampled at 5-minute intervals, were aggregated to form high-resolution glycemic trajectories spanning multiple days per subject. Data preprocessing involved handling missing values through linear interpolation for gaps shorter than 2 hours, with longer interruptions classified as non-wear time and excluded from trajectory calculations. Only 44 subjects possessed complete clinical annotations for insulin resistance (IR) and β-cell dysfunction, forming the labeled subset used for downstream modeling. These records were partitioned into training and validation sets using a stratified split to preserve the underlying class distribution. All preprocessing steps were implemented in Python, with deterministic random seeds fixed to ensure reproducibility across experimental runs.

B. Feature engineering framework

To address the high dimensionality and non-linear dynamics of raw glucose trajectories, we engineered a comprehensive set of 26 features explicitly designed to capture multi-faceted glycemic regulation. These features were systematically grouped into six physiological and computational categories: (1) basic statistics, including mean, median, standard deviation, and range of glucose levels; (2) variability indices, such as coefficient of variation (CV), mean amplitude of glycemic excursions (MAGE), and standard deviation of glucose (SDG); (3) time-in-range (TIR) metrics, quantifying the percentage of readings within the target range (70–180 mg/dL) as well as time spent in hypoglycemic (< 70 mg/dL) and hyperglycemic (> 180 mg/dL) zones; (4) trajectory parameters, derived from temporal derivatives and autocorrelation functions to capture glycemic slope, curvature, oscillation frequency, and rate of change; (5) distributional properties, including skewness, kurtosis, and percentile-based dispersion measures to characterize tail risks and asymmetry in glycemic profiles; and (6) adherence signals, computed as the ratio of active sensor readings to total monitoring duration, along with wear-time consistency indices and data completeness ratios. This projection from raw time-series data into a structured feature space reduces measurement noise while preserving clinically relevant physiological patterns, directly addressing the challenge of translating high-dimensional CGM data into actionable metabolic insights.

C. Machine learning classifiers and model architecture

We evaluated four distinct machine learning architectures to determine optimal algorithmic inductive biases for metabolic dysregulation prediction. Logistic regression served as a linear baseline, modeling the log-odds of class membership with L2 regularization to prevent overfitting. Random forest employed an ensemble of 500 unpruned decision trees, utilizing Gini impurity for node splitting and out-of-bag error estimation to mitigate variance. Gradient boosting implemented an additive model of 300 shallow decision trees (maximum depth = 5) with a learning rate of 0.1, optimized via gradient descent on a differentiable loss function (binary cross-entropy). Finally, a multilayer perceptron (MLP) was constructed with two hidden layers of 64 and 32 neurons, respectively, utilizing rectified linear unit (ReLU) activation functions, batch normalization, and dropout regularization ($p = 0.3$) to capture complex non-linear interactions between engineered features. All models were implemented using standard scientific computing libraries, with hyperparameters optimized via grid search combined with 5-fold cross-validation on the training set.

D. Experimental setup and validation protocol

The modeling pipeline was strictly designed to prevent data leakage and ensure generalizability. The 44 fully labeled records were partitioned into training (80%) and validation (20%) sets using stratified sampling. Within the training set, 5-fold cross-validation was employed to tune hyperparameters and assess model stability. Class imbalance inherent in clinical cohorts was addressed by applying class-weighted loss functions during training, ensuring that minority classes contributed proportionally to the optimization objective. Feature scaling was performed using robust standardization (median centering and interquartile range scaling) to mitigate the influence of outliers inherent in glycemic data. All experiments were conducted under deterministic random seeds, and feature importance was extracted post-hoc to verify that model decisions aligned with physiological expectations rather than spurious correlations.

E. Evaluation metrics and performance assessment

Predictive performance was rigorously assessed using task-specific metrics tailored to the binary classification objectives. For insulin resistance and β-cell dysfunction prediction, we reported accuracy, receiver operating characteristic area under the curve (ROC-AUC), and macro-averaged F1-score. Accuracy measured overall classification correctness, while ROC-AUC evaluated the model's discriminative ability across all decision thresholds. The macro-averaged F1-score was prioritized to ensure balanced performance across both positive and negative classes, preventing dominance by the majority class. Model selection was driven exclusively by validation set performance, with gradient boosting emerging as optimal for insulin resistance classification and logistic regression demonstrating superior generalization for β-cell dysfunction prediction. Statistical comparisons were conducted to verify that performance differences aligned with the distinct linear and non-linear dependencies inherent in each metabolic pathway, confirming that carefully engineered time-series glucose features paired with appropriately selected machine learning architectures provide a robust framework for early detection.

III. RESULTS

A. Feature engineering and dataset characteristics

The preprocessing pipeline successfully transformed high-frequency CGM readings into a structured representation comprising 26 engineered features. These metrics were systematically organized across six physiological and computational domains: basic statistics (mean, standard deviation, minimum, maximum, median), variability indices (glucose range, coefficient of variation, mean glycemic deviation, J-index), time-in-range metrics (time below 70 mg/dL, time in 70–180 mg/dL, time above 180 mg/dL, time above 250 mg/dL), trajectory parameters (peak glucose, time to peak, recovery rate, area under the curve, glucose slope), distributional properties (skewness, kurtosis, 10th/25th/75th/90th percentiles), and adherence indicators (rolling coefficient of variation, sensor jump count). This projection effectively reduced measurement noise while preserving clinically relevant glycodynamic patterns.

The modeling was performed on a fully annotated subset of the clinical cohort, comprising 44 subjects with complete metabolic labels. These records were partitioned using stratified sampling into a training set ($n = 27$) and a validation set ($n = 17$), preserving the underlying class distribution. For insulin resistance, the cohort exhibited a near-balanced split (class 0: 23 subjects; class 1: 21 subjects), and similarly for β-cell dysfunction (class 0: 22; class 1: 22). This balance ensured that performance metrics were not artificially inflated by majority-class dominance, allowing for a rigorous assessment of model generalization across both metabolic pathways.

B. Insulin resistance classification performance

Classification of insulin resistance revealed distinct algorithmic behaviors across the evaluated architectures. Gradient boosting achieved the highest overall performance, yielding an accuracy of 0.772 and a macro-averaged F1-score of 0.740, with precision reaching 0.867 and recall at 0.700. The receiver operating characteristic area under the curve (ROC-AUC) for this model was 0.708. Logistic regression demonstrated moderate discriminative ability, achieving an accuracy of 0.681, a macro-F1 of 0.659, and a ROC-AUC of 0.738. Random forest produced an accuracy of 0.639 and a macro-F1 of 0.623, with a ROC-AUC of 0.770. The multilayer perceptron (MLP) exhibited the lowest performance across all metrics, recording an accuracy of 0.564, a macro-F1 of 0.331, and a ROC-AUC of 0.668.

The superior performance of gradient boosting suggests that insulin resistance manifests through complex, non-linear interactions among glycemic features. The high precision indicates strong positive-class identification, while the moderate recall reflects the inherent difficulty in capturing borderline or early-stage metabolic dysregulation. The MLP's underperformance, despite the application of batch normalization and dropout regularization, is likely attributable to the limited training sample size ($n = 27$), which restricts the network's capacity to learn hierarchical representations without overfitting. The consistent ROC-AUC values across linear and tree-based models indicate that all architectures retain some baseline discriminative capacity, but gradient boosting's additive structure optimally balances bias and variance for this specific task.

C. Beta-cell dysfunction prediction results

Prediction of β-cell dysfunction demonstrated a contrasting performance landscape, with logistic regression emerging as the most effective classifier. The linear model achieved an accuracy of 0.711, a macro-F1-score of 0.709, and the

highest ROC-AUC of 0.800. Random forest yielded an accuracy of 0.664 and a macro-F1 of 0.598, with a ROC-AUC of 0.750. Gradient boosting achieved an accuracy of 0.611 and a macro-F1 of 0.602, with a ROC-AUC of 0.695. The MLP recorded an accuracy of 0.522, a macro-F1 of 0.303, and a ROC-AUC of 0.720.

The dominance of logistic regression in this task implies that β-cell dysfunction exhibits more linear relationships with the engineered CGM features. Metrics such as time-in-range, mean glucose levels, and basic variability indices likely serve as strong linear predictors of β-cell secretory capacity. The relatively high and consistent ROC-AUC values across all non-linear models (> 0.72) suggest that the underlying signal is robust enough to be captured by simpler decision boundaries, but complex non-linear interactions may introduce unnecessary variance or overfitting in small-sample regimes. The lower precision and recall of tree-based ensembles further support the hypothesis that β-cell impairment follows a more straightforward, dose-response-like relationship with aggregated glycemic exposure rather than intricate trajectory-dependent patterns.

D. Model comparison and interpretative insights

Comparing the two classification tasks reveals a clear task-dependent performance pattern. Gradient boosting optimally captures insulin resistance, which likely involves compensatory hyperinsulinemia, altered glucose clearance kinetics, and non-linear interactions between variability indices and distributional tails. In contrast, logistic regression generalizes best for β-cell dysfunction, suggesting that this pathway is more directly proportional to sustained glycemic exposure and time-in-range metrics. The differential performance underscores the importance of matching algorithmic inductive biases to underlying pathophysiology: additive tree ensembles excel at modeling non-linear physiological feedback loops, while regularized linear models generalize more effectively when metabolic dysregulation follows proportional, cumulative exposure patterns.

Furthermore, the consistent underperformance of the MLP across both tasks highlights a critical practical consideration for clinical machine learning: deep architectures require substantially larger labeled datasets to realize their representational advantages. The application of class-weighted loss functions and robust standardization mitigated some variance, but the small validation set ($n = 17$) constrained the MLP's ability to learn stable decision boundaries. These findings validate the methodological choice of prioritizing model selection based on task-specific validation performance rather than architectural preference, reinforcing the framework's adaptability to diverse metabolic endpoints.

E. Summary of findings

The experimental evaluation demonstrates that carefully engineered time-series glucose features, when paired with appropriately selected machine learning architectures, provide a robust framework for early metabolic disorder detection. Gradient boosting achieved optimal classification of insulin resistance (accuracy: 0.772, macro-F1: 0.740), leveraging non-linear feature interactions to identify complex glycodynamic patterns. Logistic regression proved most effective for β-cell dysfunction prediction (accuracy: 0.711, ROC-AUC: 0.800), capitalizing on linear relationships between sustained glycemic exposure and secretory capacity. The task-dependent performance highlights the necessity of aligning model complexity with underlying physiological mechanisms, while the consistent discriminative ability across architectures confirms that aggregated CGM metrics contain sufficient signal for actionable clinical inference. These results establish a scalable, data-driven pathway for identifying at-risk individuals prior to conventional diagnosis, supporting proactive metabolic healthcare interventions.

IV. CONCLUSIONS

Predicting insulin resistance and beta-cell dysfunction from continuous glucose monitoring data remains a significant clinical challenge due to the high dimensionality, measurement noise, and non-linear dynamics inherent in glycemic trajectories. To address this limitation, we developed a machine learning framework that systematically translates raw high-frequency glucose readings into a compact set of twenty-six engineered features. These metrics were explicitly designed to capture multi-faceted physiological regulation, encompassing basic statistics, variability indices, time-in-range distributions, temporal trajectory parameters, distributional tail properties, and sensor adherence signals. By projecting unstructured time-series data into this structured feature space, we reduced measurement noise while preserving clinically relevant glycodynamic patterns.

The proposed framework was validated on a clinical cohort comprising four hundred fifty-seven subjects, with forty-four records containing complete metabolic annotations. Raw CGM readings sampled at five-minute intervals were preprocessed using linear interpolation for short gaps and excluded for extended non-wear periods. The labeled subset

was partitioned via stratified sampling into training and validation sets to preserve class distribution balance. We evaluated four distinct machine learning architectures: logistic regression, random forest, gradient boosting, and a multilayer perceptron. All models were trained using five-fold cross-validation with class-weighted loss functions and robust standardization to mitigate outlier influence and prevent data leakage.

Our experimental evaluation revealed that predictive performance is highly task-dependent, with algorithmic success varying significantly across the two metabolic endpoints. Gradient boosting emerged as the optimal classifier for insulin resistance, achieving an accuracy of 0.772 and a macro-averaged F1-score of 0.740. In contrast, logistic regression demonstrated superior generalization for beta-cell dysfunction prediction, yielding an accuracy of 0.711 and a ROC-AUC of 0.800. The multilayer perceptron consistently underperformed across both tasks, likely due to the limited training sample size restricting its capacity to learn stable hierarchical representations.

These findings demonstrate that carefully engineered time-series glucose features, when paired with appropriately selected machine learning architectures, provide a robust and clinically actionable framework for early metabolic disorder detection. The differential performance underscores the necessity of aligning algorithmic inductive biases with underlying pathophysiology: additive tree ensembles effectively capture the complex, non-linear feedback loops characteristic of insulin resistance, while regularized linear models generalize more effectively when beta-cell impairment follows proportional, cumulative exposure patterns. Furthermore, the consistent limitations of deep architectures on small labeled cohorts highlight a critical practical consideration for clinical machine learning. Ultimately, this work establishes a scalable pathway for leveraging routine continuous glucose monitoring data to identify at-risk individuals prior to conventional diagnosis, supporting proactive and personalized metabolic healthcare interventions.