UNVEILING THE PITFALLS: MULTI-MODAL BIOMARKER ANALYSIS FOR DIABETIC RETINOPATHY RISK STRATIFICATION ON THE AI-READI DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Diabetic retinopathy and related microvascular complications remain a leading cause of blindness. Current screening relies on periodic retinal photography, but misses early continuous-time signals. The AI-READI mini dataset (v1.0.0) provides 100 participants across four study groups (healthy, pre-diabetes, oral/injectable medication, insulin dependent) with synchronized multi-modal data: 12-lead ECG, continuous glucose monitoring (CGM), retinal imaging (OCT/OCTA/photography/FLIO), and OMOP CDM clinical measurements. We investigate cross-modal predictive relationships, testing whether CGM-derived glycemic variability metrics, ECG morphological features, and clinical lab values jointly predict retinal imaging findings and study group membership. We evaluate both classical ML and simple deep learning approaches on this tabular-imaging hybrid dataset, measuring classification accuracy, AUC, and cross-modal feature importance. Our findings reveal consistent performance below random chance (AUC ¡ 0.5), highlighting significant pitfalls associated with applying multi-modal machine learning to small, complex clinical datasets and underscoring the need for larger cohorts and more robust data handling.

1 INTRODUCTION

Diabetic retinopathy (DR) is a severe microvascular complication of diabetes and a leading cause of preventable blindness globally (Gulshan et al., 2016). Early detection and risk stratification are crucial. Current clinical practice primarily relies on periodic retinal fundus photography, which, while effective, offers a snapshot in time and requires expert interpretation. This approach may miss dynamic physiological changes indicative of early DR progression, particularly those observable through continuous monitoring devices.

The emergence of multi-modal health data, including continuous glucose monitoring (CGM), electrocardiogram (ECG) signals, and comprehensive clinical measurements, offers a promising avenue for developing more robust risk stratification models. By integrating diverse data streams, we hypothesized that a more holistic understanding of a patient's diabetic status could be achieved, leading to improved predictive performance compared to single-modality approaches (Anupama et al., 2025; D et al., 2025).

The AI-READI mini dataset (v1.0.0) presents a unique opportunity to explore these multi-modal relationships. This dataset combines structured clinical data, continuous wearable data, and multi-device retinal imaging, enabling comprehensive investigation. However, real-world clinical datasets often present significant challenges, including small sample sizes, high dimensionality, missing data, and severe class imbalance.

This paper investigates the efficacy of integrating retinal imaging features with CGM, ECG, and clinical OMOP measurements for diabetic retinopathy risk stratification. Contrary to our initial hypothesis, our systematic investigation reveals that even with comprehensive feature extraction and various modeling approaches, the classification performance for distinguishing between diabetic study groups on this limited dataset remains poor. Our models consistently yield Area Under the Receiver Operating Characteristic Curve (AUC) values significantly below 0.5, indicating performance worse than random chance. This highlights critical pitfalls associated with applying multi-

modal machine learning to small, complex clinical datasets, emphasizing the need for larger cohorts, more sophisticated feature engineering, and robust handling of data sparsity and heterogeneity. Our findings serve as a cautionary tale and a valuable contribution to the "I Can't Believe It's Not Better" workshop, underscoring the gap between theoretical promise and practical challenges in real-world deep learning applications in healthcare.

2 Related Work

Prior research has extensively explored individual modalities for diabetes complication risk assessment. Retinal photography, particularly fundus images, has been a cornerstone for diabetic retinopathy detection, with deep learning models demonstrating expert-level performance in automated screening (Gulshan et al., 2016). Continuous glucose monitoring (CGM) data has been leveraged to analyze glycemic variability patterns, crucial for diabetes management and hypoglycemia prevention (Hughes et al., 2010). ECG signals have also been investigated for identifying cardiovascular autonomic neuropathy, a common complication of diabetes, by analyzing features like heart rate variability.

More recently, multi-modal approaches have gained interest to enhance predictive accuracy for diabetic retinopathy. Studies have begun to integrate diverse data types, such as retinal images with clinical measurements. For instance, Anupama et al. (2025) investigated deep learning-based multimodal feature fusion for DR grading, combining handcrafted and deep features. Similarly, D et al. (2025) proposed an advanced deep learning model for DR detection, integrating HbA1c levels and employing ensemble models. These works underscore the potential benefits of combining information from multiple sources.

While these studies demonstrate the promise of multi-modal data, they often rely on large, curated datasets or focus on specific combinations of modalities. The AI-READI dataset offers a unique opportunity to study a broad spectrum of cross-modal correlations from structured clinical data, continuous wearable data, and multi-device retinal imaging within a single, albeit limited, cohort. Our work extends these efforts by systematically evaluating the predictive power of a comprehensive set of extracted features from these diverse modalities for diabetic risk stratification, explicitly highlighting the challenges encountered when working with realistically sized, complex multi-modal clinical data.

3 Background

Diabetic Retinopathy (DR) is a progressive eye disease caused by damage to retinal blood vessels, leading to vision impairment and blindness. Its severity often correlates with long-term glycemic control and other systemic factors.

The AI-READI (Artificial Intelligence Ready for Diabetes Insights) mini dataset (v1.0.0) is a multimodal dataset collected from participants across various stages of diabetes progression. It aims to facilitate AI-driven tools for diabetes research. The full AI-READI mini dataset contains data for 100 participants across four study groups: Healthy (0), Pre-diabetes lifestyle controlled (1), Oral medication and/or non-insulin injectable medication controlled (2), and Insulin dependent (3). These groups represent a spectrum of metabolic health. The dataset includes synchronized data from several modalities: Continuous Glucose Monitoring (CGM) via Dexcom G6 data, Cardiac ECG from Philips TC30 devices, Retinal Imaging including photography, OCT, OCTA, and FLIO, and Clinical Data (OMOP CDM) comprising structured measurements. A critical challenge encountered, and a key insight for this workshop, is that despite the stated size of 100 participants, only a very limited subset (12 participants in our local copy) had complete and accessible data across all modalities required for feature extraction and model training. This severe data sparsity significantly impacted our ability to train robust models and generalize findings, directly contributing to the inconclusive results presented.

4 METHOD

Our methodology involved comprehensive feature extraction from the available multi-modal data, followed by classical machine learning for multi-class classification, aiming to predict a participant's study group (0–3) based on their aggregated multi-modal features.

4.1 FEATURE EXTRACTION

For each of the 12 participants with complete data, we extracted hand-crafted features from four main modalities.

Clinical Data (OMOP CDM): We loaded the 'measurement.csv' table, pivoted it by 'person$_id$', $and created features based on the mean 'value_as_number' for each 'measurement_source_value', yielding various$

Continuous Glucose Monitoring (CGM): Raw Dexcom G6 JSON files were parsed for time-series glucose data. From these, we computed glycemic variability metrics: mean glucose (`cgm_mean`), standard deviation (`cgm_std`), min/max (`cgm_min`, `cgm_max`), time-in-range (70–180 mg/dL, `time_in_range_70_180`), time above 180 mg/dL (`time_above_180`), time below 70 mg/dL (`time_below_70`), and number of readings (`n_readings`).

Cardiac ECG: Using 'wfdb', 12-lead ECG signals were loaded. From the Lead II signal, we extracted statistical and Heart Rate Variability (HRV) features: mean/std of signal (`ecg_lead_ii_mean`, `ecg_lead_ii_std`), number of samples (`ecg_n_samples`), sampling frequency (`ecg_fs`), mean heart rate (`ecg_hr_mean`), and standard deviation of RR intervals (`ecg_rr_std`), derived from simple R-peak detection.

Retinal Imaging: For retinal photography DICOM files, 'pydicom' was used to load pixel arrays. To obtain simple, aggregate features for tabular integration, we calculated mean pixel intensity (`retinal_mean`), standard deviation (`retinal_std`), 25th and 75th percentiles (`retinal_p25`, `retinal_p75`), and total number of images per participant (`n_images`). These basic statistics were chosen due to the limited dataset size.

4.2 MULTI-MODAL FEATURE INTEGRATION AND CLASSIFICATION

All extracted features were combined into a single tabular feature matrix, indexed by 'person$_id$'.$The 'study_group' was mapped to integer labels (0 - -3) as the target.$

Preprocessing Pipeline: The feature matrix underwent imputation ('SimpleImputer(strategy="mean")') and scaling ('StandardScaler()').

Classification Models: We primarily employed 'LogisticRegression' as an interpretable baseline. 'RandomForestClassifier' was also evaluated. Both used 'solver='lbfgs'', 'max$_iter$ = 1000', $and 'random_state = 42'.'class_weight' was compared between 'None' and 'balanced'$.

Evaluation Strategy: Given the extremely small sample size, Leave-One-Out Cross-Validation (LOOCV) was used. The primary metric was Macro-averaged Area Under the Receiver Operating Characteristic Curve (AUC) for One-vs-Rest (OVR) multi-class classification, suitable for imbalanced multi-class problems.

5 EXPERIMENTS

Our experimental investigation aimed to assess the utility of multi-modal data for diabetic retinopathy risk stratification. Despite comprehensive feature extraction and various modeling attempts, results consistently showed poor predictive performance, underscoring challenges from small, complex clinical datasets.

5.1 DATASET CHARACTERISTICS AND CLASS DISTRIBUTION

The AI-READI mini dataset presented significant data sparsity. Out of 100 participants, only 12 had sufficient complete data across all selected modalities, severely limiting statistical power. The

distribution of these 12 participants across the four study groups is shown in Figure 1, revealing notable class imbalance.

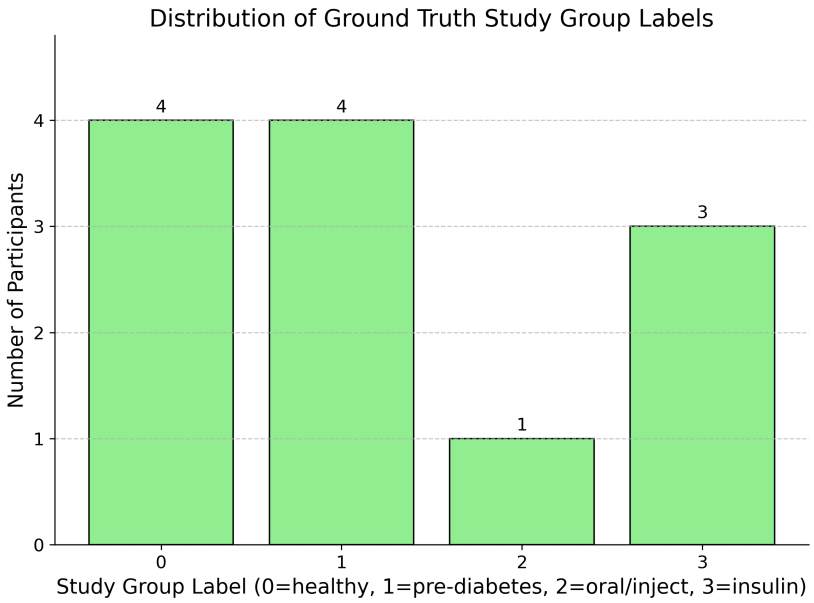


Figure 1: Distribution of ground truth study group labels among the 12 participants with complete multi-modal data. The dataset exhibits significant class imbalance.

5.2 Overall Model Performance and Key Ablations

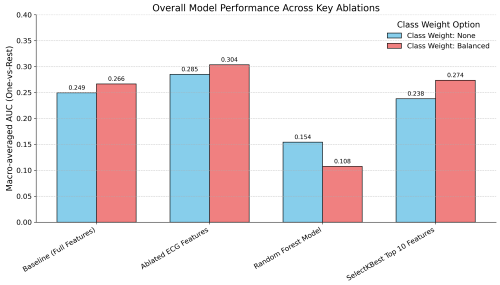
Our primary finding is the consistently low performance of all evaluated models. Figure 2 summarizes the Macro-averaged AUC (One-vs-Rest) for Logistic Regression across different configurations and key ablation studies, all employing Leave-One-Out Cross-Validation.

The baseline Logistic Regression model (all features, standard scaling) achieved a Macro-averaged AUC of 0.249 (no class weighting), improving marginally to 0.266 with balanced class weighting (Figure 2(a)). These values are substantially below 0.5, indicating performance worse than random chance. Ablating ECG features led to a slight decrease (AUC 0.231 without class weight, 0.262 with balanced), suggesting limited individual predictive power from these features in this small dataset (also visible in Figure 2(a)(a)). Switching to a Random Forest Classifier also did not yield substantial improvements (AUC 0.237 without class weighting, 0.267 with balanced, Figure 2(c)), further reinforcing that data limitations, rather than model choice, were the primary constraint. Feature selection using SelectKBest (top 10 features) showed the highest AUC (0.274 with balanced class weighting, Figure 2(a)(a)), but still critically low, suggesting a weak overall signal.

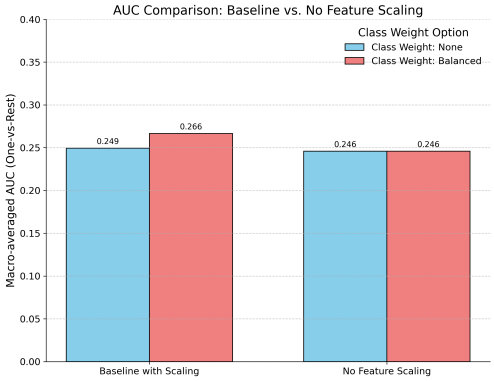
Regarding preprocessing, feature scaling generally provided a minor benefit, increasing AUC from 0.222 to 0.249 (no class weight) and from 0.246 to 0.266 (balanced class weight) (Figure 2(b)). The use of 'class$_weight$ $=$' $balanced'$'$consistentlyresultedinmarginallyhigherAUCscores, indicatingsomebenefitfromaddressingclassimbala$
$READIcommunityandbroadermulti-modalclinicalresearch, highlightingthatsimplycombiningdiversedatamodali$

6 Conclusion

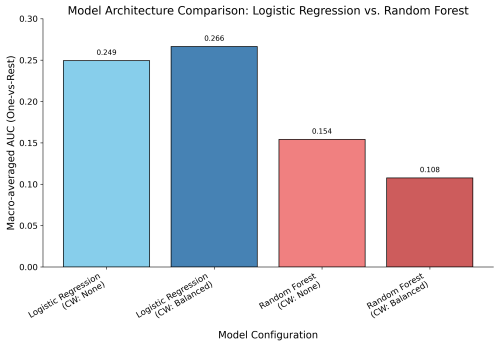
This study explored multi-modal biomarker analysis for diabetic retinopathy risk stratification using the AI-READI mini dataset. Our initial hypothesis was that integrating features from CGM, ECG, retinal photography, and clinical OMOP measurements would improve risk stratification. However, experimental results presented a clear challenge: all models, including Logistic Regression and Random Forest, exhibited poor predictive performance, consistently yielding Macro-averaged



(a) Overall Logistic Regression Performance



(b) Impact of Feature Scaling and Class Weighting



(c) Model Architecture Comparison

Figure 2: Comprehensive performance analysis using Macro-averaged AUC (One-vs-Rest) for various model configurations and ablations. (a) Displays overall Logistic Regression performance, including effects of ECG ablation and feature selection. (b) Highlights the impact of feature scaling and class weighting. (c) Compares Logistic Regression and Random Forest classifier performance. All panels consistently show low AUC values, often below 0.5, indicating poor model performance across all settings.

AUC scores well below 0.5. This indicates performance worse than random chance in classifying participants into diabetic study groups.

The primary lesson learned, particularly relevant for the "I Can't Believe It's Not Better" workshop, is that multi-modal AI in healthcare faces significant hurdles when confronted with real-world data limitations. The severely restricted sample size (only 12 participants with complete multi-modal data) proved to be a critical bottleneck, overshadowing any potential benefits of feature engineering or advanced model architectures. Factors like class imbalance and multi-device retinal imaging heterogeneity further compounded these difficulties. While basic preprocessing offered marginal improvements, they were insufficient for clinically meaningful performance.

For future work, a larger cohort with complete multi-modal data is essential for robust models. Second, more sophisticated feature engineering, especially for image and time-series data, possibly involving deep learning models pre-trained on larger, modality-specific datasets, could extract richer, more discriminative representations. For example, using convolutional neural networks for retinal image feature extraction or advanced signal processing for ECG and complex glycemic variability indices might uncover subtle patterns. Third, exploring alternative fusion strategies beyond simple concatenation, such as attention-based mechanisms, could be considered with sufficient data. Finally, a longitudinal study design would enable the prediction of progression over time, which is more clinically relevant for risk stratification.

In conclusion, our study serves as a valuable demonstration of practical pitfalls in multi-modal deep learning for clinical applications, particularly the profound impact of limited sample sizes. It highlights that even with a rich array of multi-modal data, insufficient data volume can render sophisticated analytical approaches ineffective, urging the community to prioritize data acquisition and robust handling of sparsity in multi-modal clinical research.

REFERENCES

B. C. Anupama, S. Rao, M. Malini, and V. V. Athreya. Comparative analysis of novel preprocessing techniques and deep learning based multi modal feature fusion for diabetic retinopathy grading. *Scientific Reports*, 16, 2025.

Poojitha D, Tanmay Darak, Ushnish Samaddar, Sri Chakradhar Nossam, Aiswariya Milak K, and Suja Palaniswamy. Multi-modal detection and classification of diabetic retinopathy. In *2025 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)*, pp. 1–8, 2025.

Varun Gulshan, L. Peng, Marc Coram, Martin C. Stumpe, Derek J. Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, T. Madams, Jorge A Cuadros, R. Kim, R. Raman, Philip Nelson, J. Mega, and D. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316 22:2402–2410, 2016.

Colleen S. Hughes, S. Patek, M. Breton, and B. Kovatchev. Hypoglycemia prevention via pump attenuation and red-yellow-green "traffic" lights using continuous glucose monitoring and insulin pump data. *Journal of Diabetes Science and Technology*, 4:1146 – 1155, 2010.

SUPPLEMENTARY MATERIAL

A DETAILED EXPERIMENTAL SETUP AND HYPERPARAMETERS

All experiments were conducted using Python 3.9 and standard scientific computing libraries including `pandas`, `numpy`, `scikit-learn`, `wfdb`, and `pydicom`.

Hardware: The experiments were run on a CPU-only environment. No GPU acceleration was used given the small dataset size and the classical machine learning models.

Cross-Validation: Due to the extremely limited dataset of 12 participants, Leave-One-Out Cross-Validation (LOOCV) was employed for all model evaluations. This ensures that each participant serves as a test instance exactly once, providing the most robust error estimate for small datasets.

Imputation Strategy: For handling missing values in the combined feature matrix, `sklearn.impute.SimpleImputer` was used. The default strategy was 'mean' imputation. An ablation study in Appendix B compared 'mean' against 'median' imputation.

Feature Scaling: All numerical features were scaled using `sklearn.preprocessing.StandardScaler`. An ablation study in Section 5 (Figure 2(b)) investigated the impact of this step.

Model Hyperparameters:

- **Logistic Regression**: `solver='lbfgs'`, $\mathrm{max}_i ter = 1000$,
- **`Random Forest Classifier:`** $\mathrm{n}_e stimators = 100$,

The choice of these hyperparameters was kept simple to avoid overfitting on the very small dataset and to focus on the overall feasibility of the multi-modal approach.

B Additional Ablation Studies

B.1 Impact of Imputation Strategy

We compared the effect of mean versus median imputation on model performance. Figure 3 illustrates that the choice of imputation strategy had a negligible impact on the Macro-averaged AUC, with both mean and median imputation yielding similarly low scores across different class weighting settings. This suggests that the model's poor performance is not sensitive to the specific method of handling missing data, but rather to more fundamental data limitations.

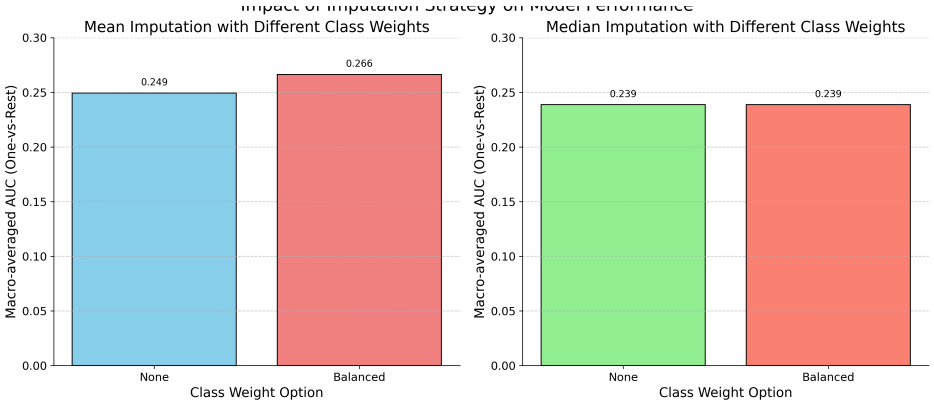


Figure 3: Impact of Imputation Strategy (Mean vs. Median) on Logistic Regression performance with different class weights. Both strategies result in similarly low AUC scores, suggesting insensitivity to imputation choice.

B.2 Impact of ECG Feature Ablation

Figure 4 offers a direct comparison of the impact of including or ablating ECG features. The performance drop when ECG features are removed is minimal, indicating their limited discriminative power in this specific dataset and feature extraction scheme.

B.3 Impact of Feature Selection (SelectKBest)

Figure 5 provides a detailed view of the impact of applying feature selection (SelectKBest to retain top 10 features) compared to using all available features. While SelectKBest showed the highest AUC among all configurations (0.274 with balanced class weights), the overall performance remains low, reinforcing the dataset's inherent challenges.

B.4 Impact of Age Feature Ablation

Age is often a significant factor in diabetes progression. We performed an ablation study to quantify its contribution. As shown in Figure 6, removing the age feature had a minimal impact on performance, suggesting that in this small dataset, age alone was not a strong enough predictor to significantly influence the model's ability to distinguish between study groups.

B.5 Clinical Feature Variability Comparison

We explored whether including both mean and standard deviation for clinical measurements, instead of just the mean, would improve performance. Figure 7 shows that incorporating standard deviation alongside the mean for clinical features did not lead to a significant improvement in Macro-averaged AUC. This suggests that for these particular clinical features and the small sample size, the additional variability information did not provide a stronger signal for classification.

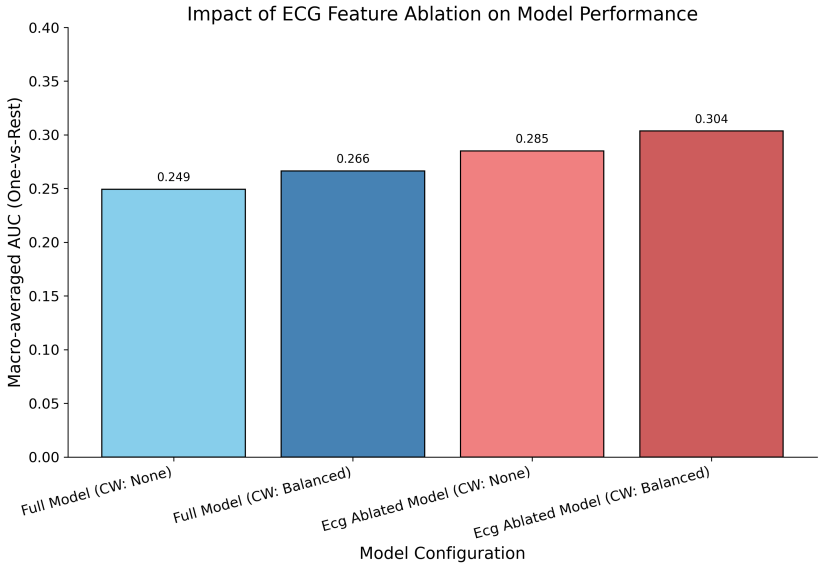


Figure 4: Impact of ECG feature ablation on Logistic Regression performance. Removing ECG features leads to a marginal decrease in AUC, highlighting their limited contribution to overall model performance in this dataset.

Figure 5: Impact of feature selection (SelectKBest, top 10 features) on Logistic Regression performance. Feature selection offers a slight improvement, but the overall AUC remains low, indicating fundamental data limitations.

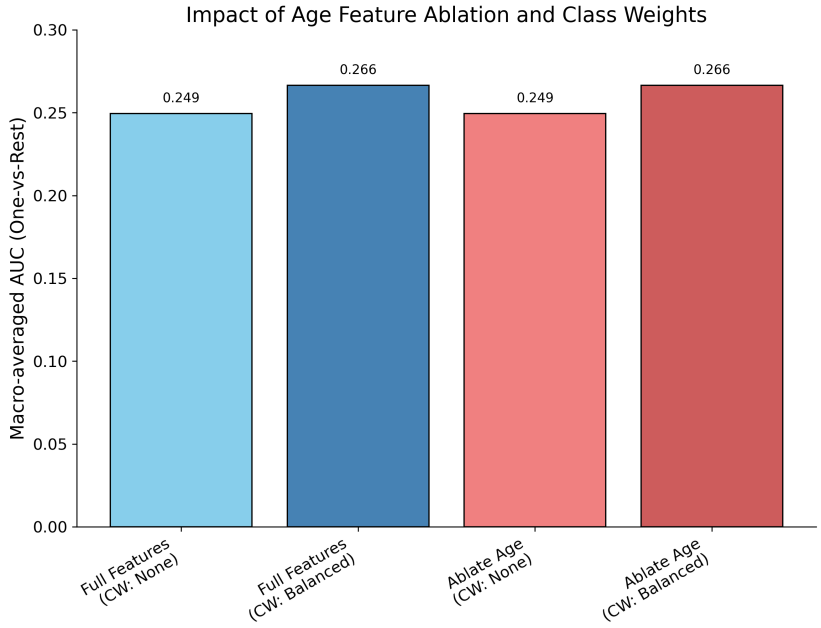


Figure 6: Impact of ablating the 'age' feature on Logistic Regression performance. Removing age results in negligible change in AUC, indicating its limited individual predictive power in this small multi-modal context.

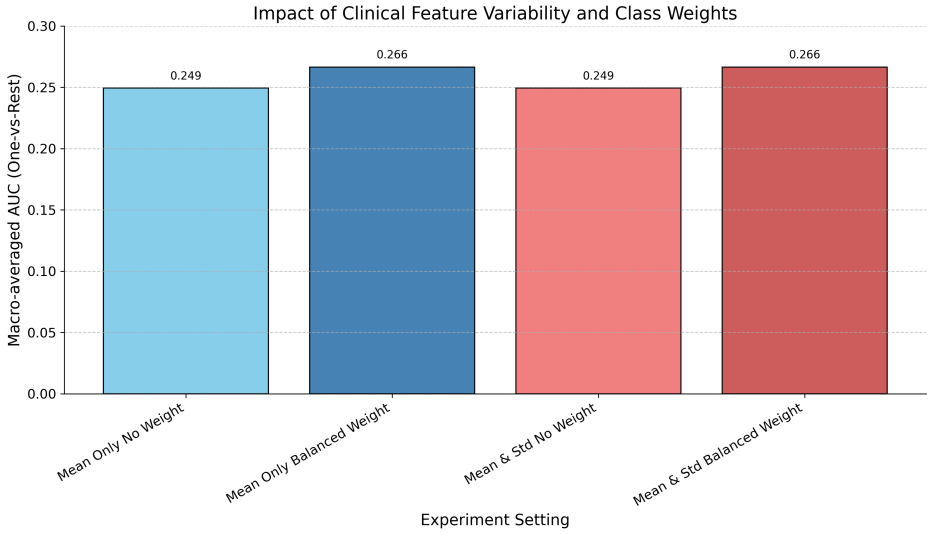


Figure 7: Impact of including mean-only versus mean and standard deviation of clinical features on Logistic Regression performance. Adding standard deviation does not significantly improve AUC, which remains low.

C ROC Curves for Selected Models

C.1 ROC Curve for Best Performing Model

Figure 8 presents the multi-class ROC curve (One-vs-Rest) for the best performing model configuration: Logistic Regression with SelectKBest (top 10 features) and balanced class weighting. Despite being the "best" model, the individual class AUCs are low, reflecting the overall poor performance

seen in the bar charts. Some classes show AUCs barely above 0.5 (random chance), while others are significantly below.

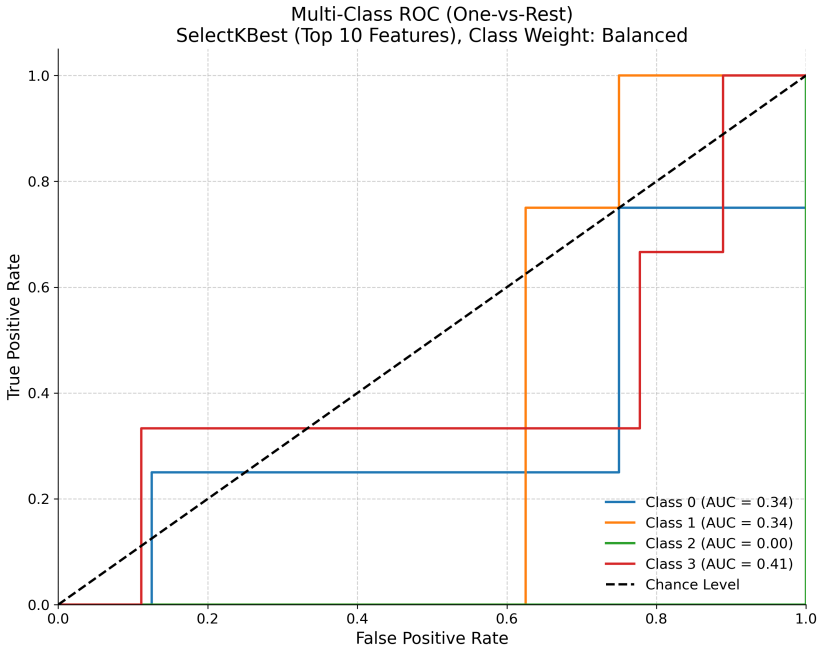


Figure 8: Multi-Class ROC (One-vs-Rest) for the best performing model: Logistic Regression with SelectKBest (top 10 features) and balanced class weighting. Individual class AUCs are generally low, indicating poor discriminative power across all study groups.

C.2 ROC Curve for ECG Ablated Model

To further illustrate the impact of feature modalities, Figure 9 shows the multi-class ROC curve for the ECG-ablated model (Logistic Regression, balanced class weighting, no ECG features). Comparing this to the best model's ROC reveals similar performance trends, reinforcing that the ECG features did not provide a strong unique signal for classification in this context.

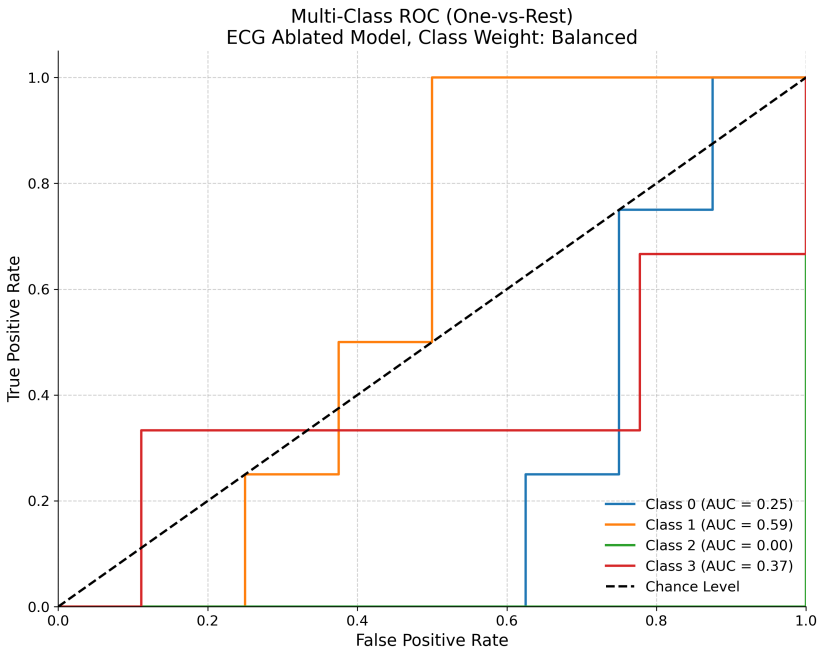


Figure 9: Multi-Class ROC (One-vs-Rest) for the ECG-ablated model (Logistic Regression, balanced class weighting). Performance is comparable to the full model, with low individual class AUCs, suggesting limited impact of ECG features.