Analysis Plan

I will begin by preprocessing – maybe consider dimensionality reduction techniques and removing features of high collinearity - data from the AI-READI dataset to align continuous glucose monitoring (CGM) values with time-series data collected from wearable sensors. This will include time-varying features such as heart rate, physical activity, respiratory rate, stress levels, sleep patterns, and environmental exposures (e.g., temperature, humidity, particulate matter, and light levels). These features will be merged with CGM readings into structured sequences for prediction. I will also incorporate static covariates—such as age, weight, questionnaire responses on diet and substance use, and social determinants of health—to capture inter-individual variability. I will use the pre-identified training, validation, and test splits provided by the dataset to ensure fair and reproducible model evaluation.

My primary forecasting models will be Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, which are well-suited for sequential physiological data. I will train these models to predict both continuous blood glucose values and binary outcomes (such as hyperglycemia and hypoglycemia) -add threshold- at increasing forecast horizons: 15 minutes, 30 minutes, 1 hour, 2 hours, 4 hours, and 8 hours. To benchmark the performance of deep learning models, I will also fit more traditional and machine learning models, including Autoregressive Integrated Moving Average (ARIMA) models and gradient-boosted decision trees using XGBoost. For these, I will engineer time-series features such as lagged glucose values, temporal indicators, and rolling averages - expound.

I will conduct hyperparameter tuning on each model type using the validation split – rewrite: "Each model will be tuned to determine optimal hyperparameters using the validation set. Write out or visually show approach for hyperparameter tuning – option 1: perform cross-validation grid search using only train data; option 2: join train and valid sets together and then perform CV grid search (also ask in office hours) For RNNs, I will tune parameters including the number of hidden units, sequence length, learning rate, dropout rate, and number of layers. For XGBoost, I will tune tree depth, learning rate, number of estimators, and regularization terms. Model performance will be evaluated using root mean square error (RMSE) and mean absolute error (MAE) for continuous predictions, and area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and precision-recall metrics for binary outcomes. I will also stratify performance results by subgroups defined by diabetes status, sex, disease severity, and social determinants of health, to investigate patterns of differential model performance and potential disparities.

To gain insight into the factors driving model predictions, I will apply interpretable machine learning techniques. For XGBoost models, I will use SHAP (SHapley Additive exPlanations) values to identify influential features at both global and individual levels. For deep learning models – will SHAP be used?, I will use methods more appropriate for complex temporal structures, including Integrated Gradients, feature ablation, and attention mechanisms if incorporated into the architecture. These interpretability tools will help me understand which physiological, behavioral, and environmental signals are most predictive of adverse glycemic events, and may inform the design of personalized interventions and future modeling work.