Using Machine Learning to Predict HbA1c and Analyze Glucose Management Strategies in Pediatric Type 1 Diabetes

Nicolas Fertout — Department of Statistics, Stanford University
Thomas Bougnon — Department of Aeronautics, Stanford University

Introduction

Context: Study conducted by the Stanford Diabetes Research Center on pediatric type-1 diabetes. Patients with type-1 diabetes wore Continuous Glucose Monitoring (CGM) devices collecting real-time glucose data every 5 minutes for one year.

Simultaneously, multiple hemoglobin A1c (HbA1c) measurements were taken to assess average blood sugar levels.

Goal: Leverage supervised and unsupervised learning techniques to predict HbA1c levels using CGM data and patient-specific information.

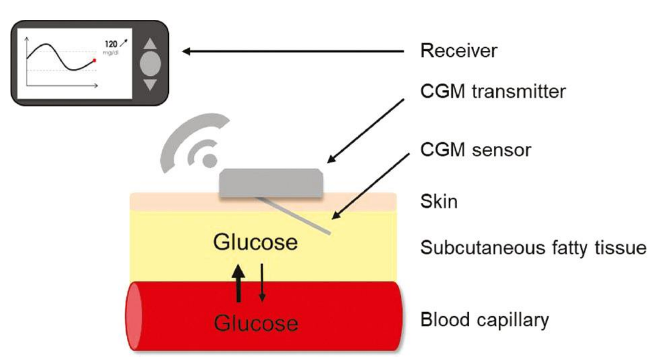


Figure 1: Scheme of CGM

Data

We selected publicly available data from the study [2].

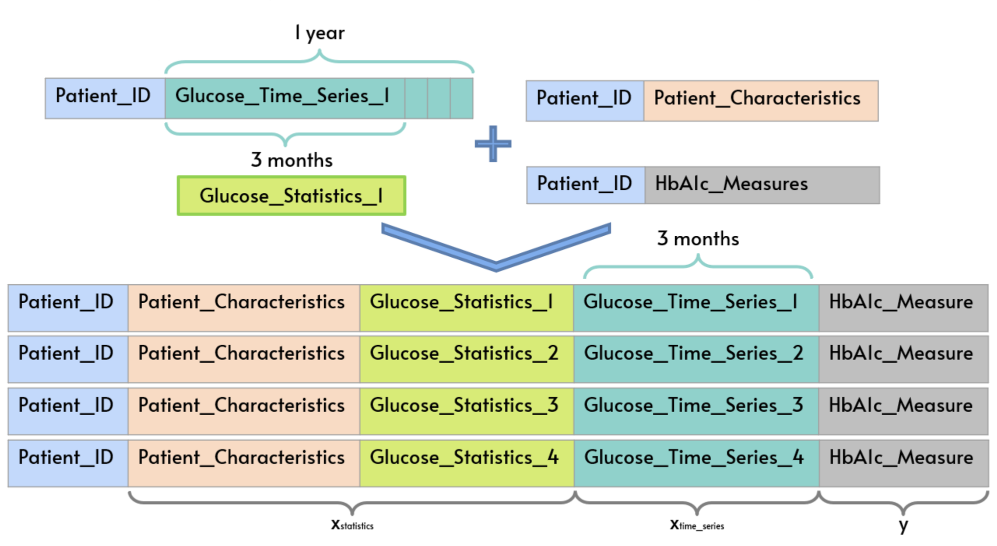


Figure 2: Hand-crafted Dataset structure

Feature Engineering: We replicated and enhanced CGM statistics used in a previous paper [3], including Mean, SD, CV, and time spent in various glycemic ranges. We also introduced new statistics like Min, Max, Median, MAD, and ROC-related metrics.

Models

Linear Models

- **Linear Regression (OLS):** baseline model to assess achievable performance with minimal effort and compare our results to the original paper.
- **Ridge Regression:** to handle multicollinearity and prevent overfitting by introducing an L2 penalty to the LS objective. It helps improve the robustness by shrinking the coefficients.
- **Lasso Regression:** to perform feature selection by adding an L1 penalty term to the LS objective. It allowed us to identify and prioritize important features, leading to a more interpretable model: only 12 input variables are kept.

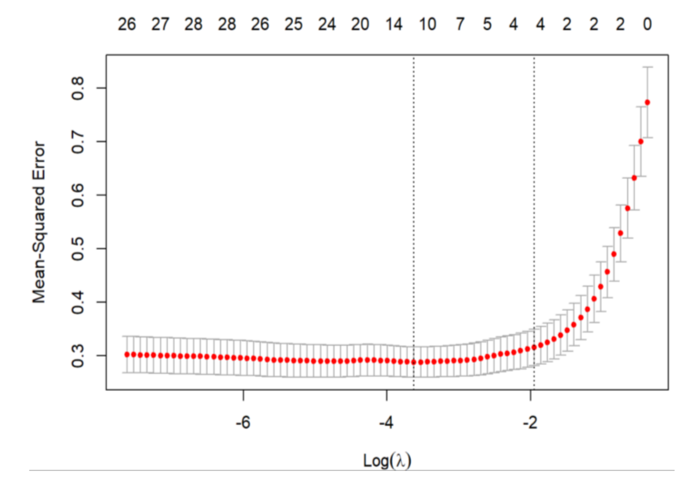


Figure 3: Cross-Validation Plot for Lasso

Non-Linear Models

- **Random Forest:** an ensemble method that builds a collection of decision trees that work independently and reduce correlation between predictions.
- **AdaBoost:** sequentially trains weak learners and assign higher weights to samples with higher regression error.
- **Gradient Boosting:** sequentially trains weak trees with each subsequent tree aiming to correct the mistakes made by the previous ones.
- **XGBoost:** optimized implementation of gradient boosting, to enhance model performance through parallel computing and regularization techniques.

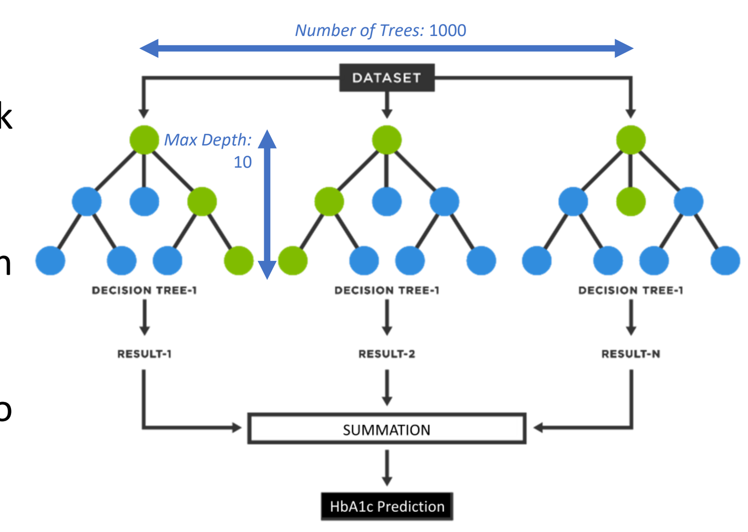


Figure 4: XGBoost Model

Deep Learning Models

- **1D-CNN:** uses a mix of convolutional and fully connected layers, effective at capturing short-term dependencies in sequential data.
- **LSTM:** Recurrent Neural Network model, effective at capturing long-term dependencies in sequential data.
- **Auto-Encoder (with RF):** uses the latent representation of CGM time series combined with patient characteristics as input for a Random Forest.
- **Transfer Learning (with RF):** benefits from pre-trained 1D-CNN to extract features, combined with patient characteristics as input for a Random Forest.

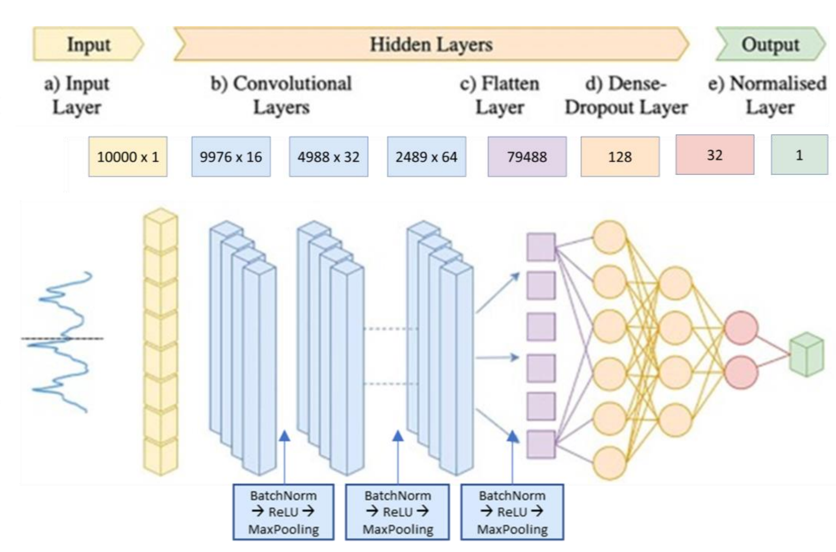


Figure 5: 1D-CNN Structure

Results

Using Hand-Crafted Features

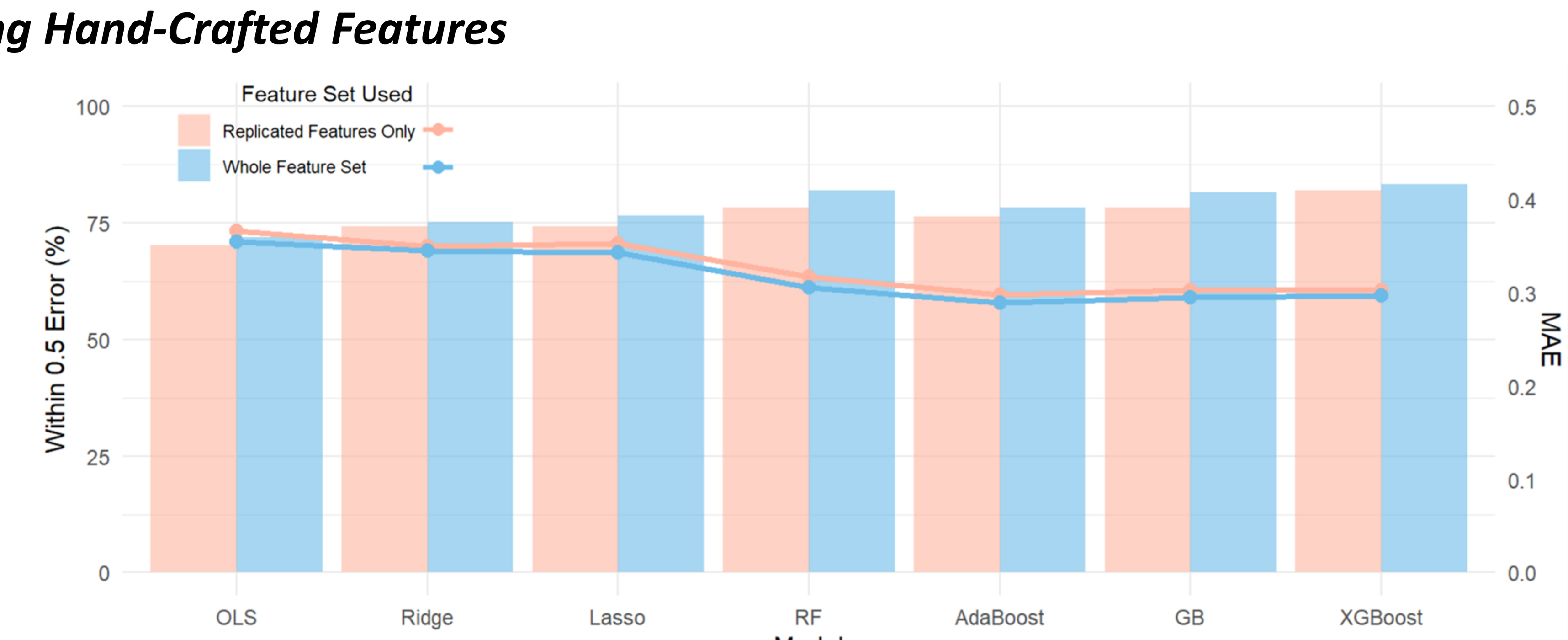
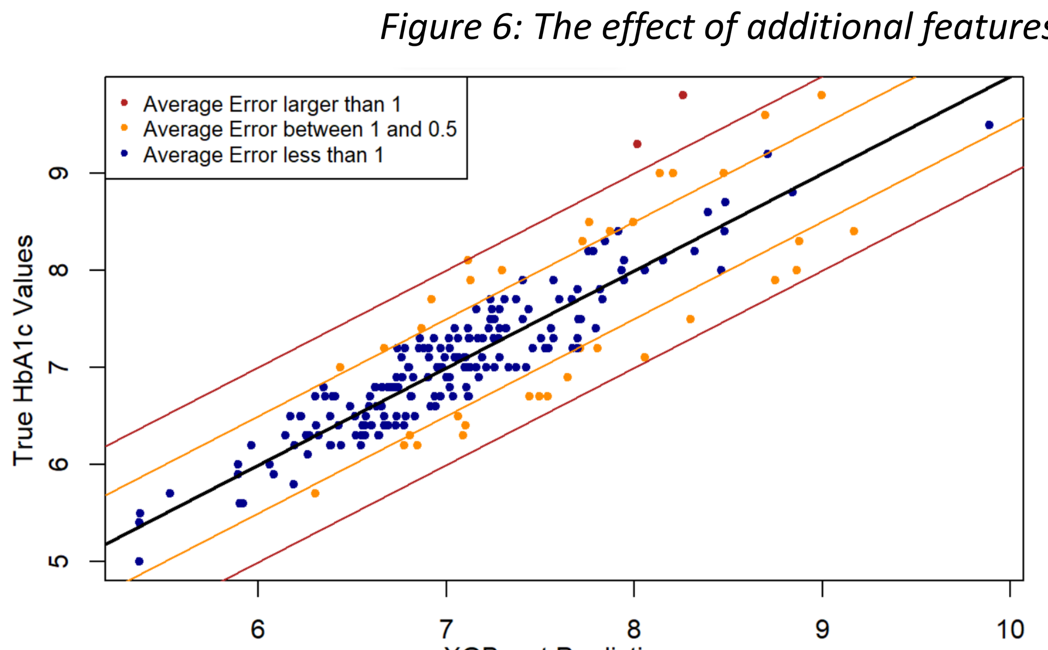


Figure 6: The effect of additional features on various model performances



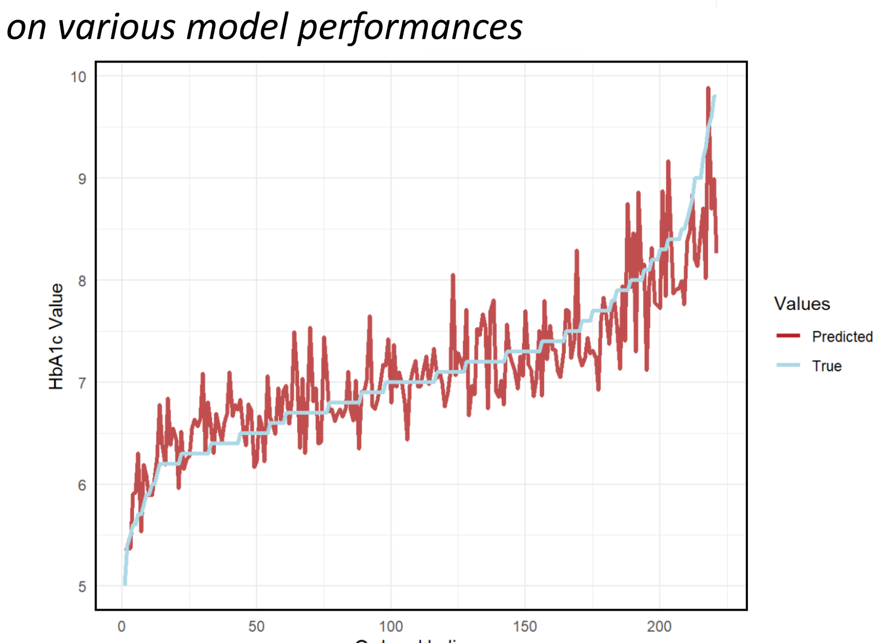


Figure 7 and 8: True HbA1c against Prediction made by XGBoost

Using Raw Time Series

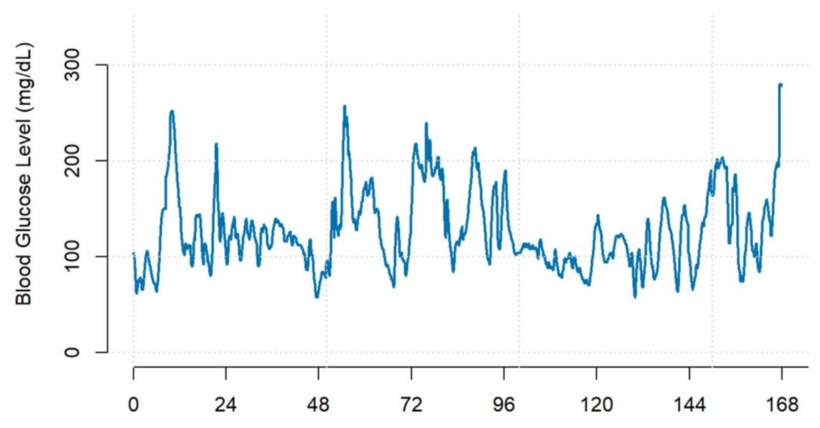


Figure 9: CGM Time Series – 1 Week

Model	Dataset	MAE
Null (Intercept Only) Model	CGM and HbA1c	0.61
1D-CNN	CGM and HbA1c	0.55
LSTM	CGM and HbA1c	0.59
Transfer Learning (with RF)	CGM and HbA1c	0.85
Auto-Encoder (with RF)	CGM and HbA1c	0.91
Null (Intercept Only) Model	Simulated ECG	1.49
1D-CNN	Simulated ECG	0.12
LSTM	Simulated ECG	0.15

Table 1: Deep Learning Model Performances

Discussion

- Overall, **additional features improved predictive power** of all models.
- **The best performing Linear Model is Lasso,** achieving an MAE of 0.343, with high accuracy. It is worth mentioning that the **optimal Lasso uses 12/30 features**, while a λ one SE away from the optimal λ gave close results using only 4/30 variables: *Mean Glucose, Proportion in Target, Proportion Far Above Target, Height*.
- **The best performing model is XGBoost** in terms of Within 1 and 0.5 Percent Point Accuracy (99.10% and 83.26% resp.), even though **AdaBoost gives the smallest MAE** (0.289). Both **outperformed RF**, which was the best-performing model in the original paper [3].
- **All Deep Learning approaches underperformed**. When **running the same models on a simulated ECG dataset of large sample size** to predict Heart Rate, yields **very good performance**. Thus, the poor performance of DL methods seems to be caused by a too small sample size, but most importantly a **lack of structure in the CGM data**, as our model worked well on the very structured ECG data and managed to still get reasonable predictions for smaller sample sizes.
- In conclusion, **this project demonstrated the effectiveness of machine learning techniques in predicting HbA1c levels** in pediatric type-1 diabetes.

Future Steps

- Identify the main differences in patient glucose management based on when/if the patient started using an insulin pump and a closed-loop system.
- Learn indicators that can predict when to recommend support from the care team or when the care team should be alerted about a patient's condition. Compare the results of those indicators with the GRI (Glycemia Risk index) [4] to see if this new metric can accurately predict patient condition.

References

[1] Overview: A New Technology-Enabled Care Model for Pediatric Type 1 Diabetes
[2] Dataset: Continuous glucose monitoring and intensive treatment of type 1 diabetes
[3] Original Paper: Improved individual and population-level HbA1c estimation using CGM data and patient characteristics
[4] Next Steps: A Glycemia Risk Index (GRI) of Hypoglycemia and Hyperglycemia for Continuous Glucose Monitoring Validated by Clinician Ratings