Assigned:		Due :
September 27, 2020	Project - Proposal & Outline	October 11, 2020

Group Members: Emily Tao <etao@hawk.iit.edu> (leader)
Yin Yang <yyang191@hawk.iit.edu>
Jim Witkiewicz <jwitkiew@hawk.iit.edu>
Geongu Park <gpark7@hawk.iit.edu>
Woojin Choi <wchoi6@hawk.iit.edu>

❖Project Proposal: Tidepool Diabetes Data

➢ A formal description of the project with a stated research goal.

Blood glucose monitoring is essential for diabetes management. With modern technology, patients with diabetes can continually monitor their blood glucose levels and adjust insulin doses, striving to keep blood glucose levels as close to normal as possible. Blood glucose levels that deviate from the normal range can lead to serious short-term and long-term complications. An automatic prediction model that warns people of imminent changes in their blood glucose levels would enable them to take preventive action. Further, automatic detection can enable automatic insulin delivery, lifting the burden from the patient to continually manage their health.

In this project, we will delve into predicting a subject's blood glucose level by applying univariate/multivariate time series models and some deep learning techniques on CGM (Continuous glucose monitoring) data. We will utilize the Tidepool diabetes data, which is a time series data containing the data from blood glucose monitor users who are Type 1 diabetic.

➢ A specific question or set of questions that the project seeks to address.

- Can we predict exact blood glucose values at least one step ahead? (1 step = 5 minutes)
- How many steps ahead can we predict accurately?
- Can we predict the trend (up/down) of the blood glucose values?

➢ A proposed methodology/approach to the analysis that will be performed.

- Steps-ahead prediction of exact values:
  - Autoregressive integrated moving average (ARIMA) forecasting.
  - Long Short-Term Memory (LSTM) forecasting.
  - Both univariate and multivariate approaches will be used.
- Trend of blood glucose values:
  - Logistic Regression incorporating multivariate features to predict whether blood glucose increases or decreases from a previous time point.

➢ A metric or set of metrics which will measure analysis results.

- Steps-ahead prediction of exact values (regression):

- Akaike's Information Criterion (AIC) :
  Useful in selecting predictors for regression, the AIC is also useful for determining the order of an ARIMA model.
- Quantile-Quantile Plot:
  The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. We can use this to analyze the residuals of our model results.
- MSE, RMSE:
  Overall error that can be used to fit and compare models.
- Plot the forecast trend line for visual confirmation

■ Trend of blood glucose values (classification):
- Confusion matrix for accuracy, precision, recall analysis.
- ROC analysis

❖ Project Outline

➢ Literature review and related work

■ Sean T. Doherty & Stephen P. Greaves (2015) Time-Series Analysis of Continuously Monitored Blood Glucose: The Impacts of Geographic and Daily Lifestyle Factors. *Hindawi*, 2015.

- Very similar dataset in variables captured & overall analysis. Notably, patients managed their food and insulin intake successfully causing these factors not to influence their blood glucose over time.

■ Q. Sun, M. V. Jankovic, L. Bally and S. G. Mougiakakou, Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network. 14th Symposium on Neural Networks and Applications (NEUREL), Belgrade, 2018, pp. 1-5, doi: 10.1109/NEUREL.2018.8586990.

- This article focuses on comparing the accuracy of the results of EMD-LSTM and LSTM models in predicting blood glucose values. The data collection is similar to our data set. We may use the unit structure of LSTM as a reference.

■ Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020.

➢ All data sources and reference data with descriptions

In total there are 300 data files in the Tidepool data set, each corresponding to an individual user's CGM data. Because the data set is so large, we will choose one representative data file to prepare our analysis, with the intention that the analysis can be extended to other data files without loss of performance. Our candidate file is titled "S_SET03_CASE005_M2_P2" and contains the following

variables:

Variable name	*Description (thanks to Reza Askari, Cinar Lab at IIT)*	*# observations*
Time	A continuous time-series variable generated by converting date-time to epoch time. This variable has been imputed and freed from outliers.	20,763 obs.
IMPUTED	CGM values which are imputed by using cubic spline (up to 50 samples) and its outliers are also removed from the variable. This is a continuous time series variable.	20,179 obs. (584 missing)
Denoised	Same CGM values with some smoothing and signal processing.	19,995 obs. (768 missing)
DAY_WEEK	A discrete variable (with integer values from 1 to 7) determining the day of the week that the data has been recorded.	20,763 obs.
YEAR, MONTH, DAY, HOUR, MIN, SEC	Extracted time-date variables suitable for classification purposes.	20,763 obs.
BASAL	A continuous variable determining the rate of basal insulin is being injected. This variable is also imputed.	20,763 obs.
bolus*	The amount of bolus insulin administered after carbohydrate intake.	473 obs.
Activity_Name*	A categorical variable (an integer ranging from 1 to 40) determining the type of physical activity.	71 obs.
Distance_value*	The distance the patient walked (in miles).	71 obs.
Activity_Duration*	A sparse variable determining the duration of physical activity.	71 obs.
insulinOnBoard*	The calculated insulin on board values for suggesting insulin to be injected.	387 obs.
insulinSensitivity*	Insulin sensitivity values.	387 obs.
smbg_new*	Self-monitoring blood glucose concentration.	876 obs.
nutrition_carbohydrate_net*	The carbohydrates entered in a healthkit food entry.	0 obs.
carb_input_new*	The amount of consumed carbohydrate (in gram)	387 obs.

Features with an asterisk (*) are sparse, meaning they contain mostly NaN values with a few scattered numeric values. IMPUTED and Denoised, corresponding to the actual blood glucose values, are not sparse but may contain runs of >50 missing samples.

➢ Data processing and pipeline

- Treatment of missing values:
  - **bolus**, **carb_input_new**:
    No value indicates no bolus insulin or carbs were taken, while positive numbers indicate the amount of bolus/carbs taken. Thus we may convert all NaN to 0.

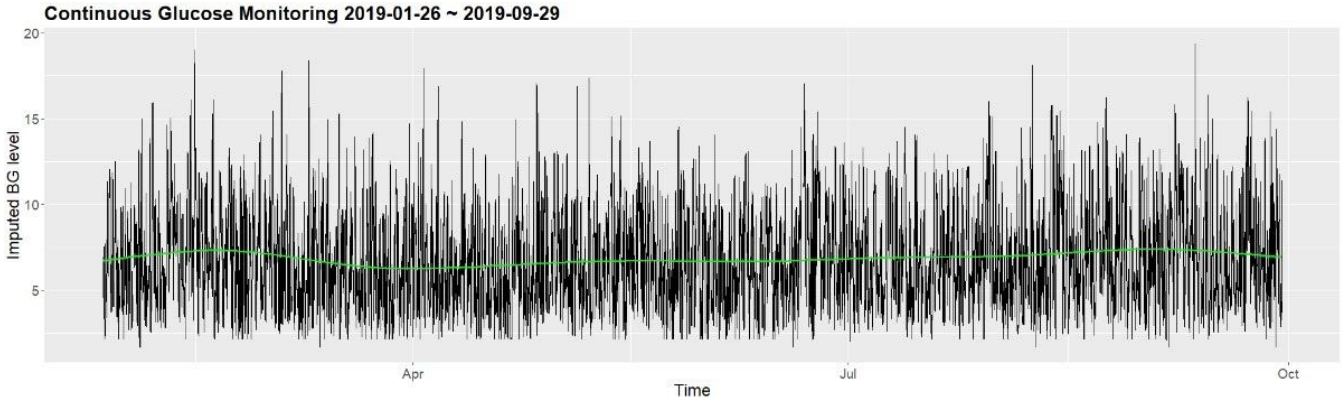
- **Activity_Name**:
  This is a categorical variable. We can use one-hot encoding for each of the values 1-40, where NaN values will simply have 0 for all dummy variables.
- **Distance_value**, **Activity_Duration**:
  If no value, then we assume no distance was walked or no activity was taken (NaN=0).
- **insulinOnBoard**, **insulinSensitivity**, **nutrition_carbohydrate_net**:
  These variables appear to be highly correlated with bolus delivery. For clarity, these variables may be dropped from analysis.
- **smbg_new**:
  These values are correlated with the target (imputed CGM) values, being sparse user-prompted blood glucose measurements at select time points. The timing often coincides with bolus delivery.

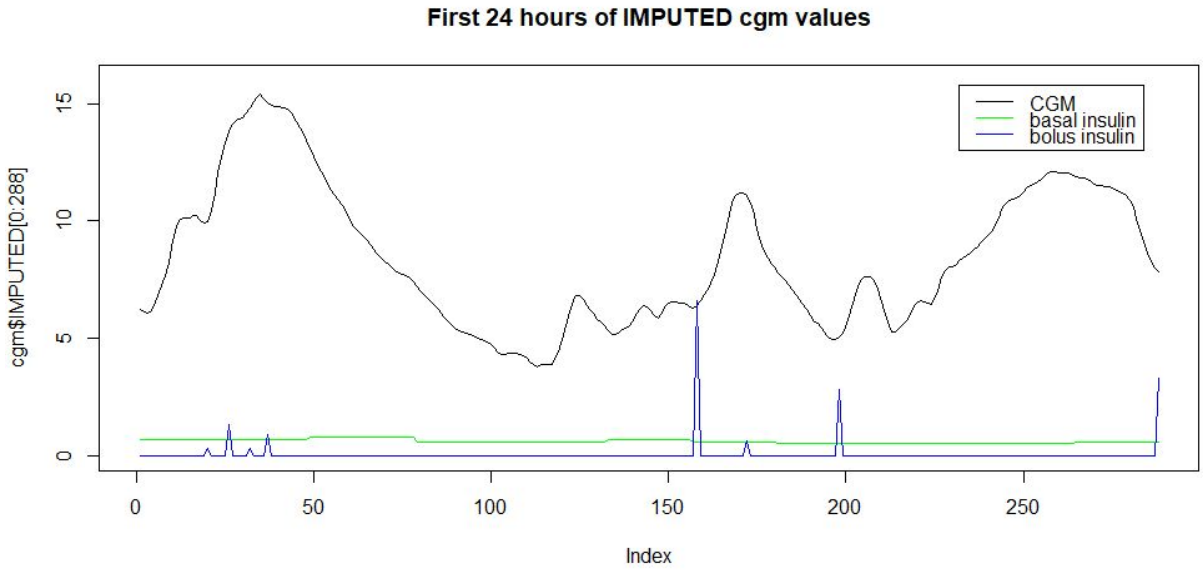
■ Preparation for analysis methods:

- Some methods such as deep learning may require the data to be reformatted into a different shape, for example [samples x timepoints x features].
- For logistic regression/classification, the target variable must be created for the change in blood glucose value from a previous time point, e.g. 1=positive change, 0=negative change.

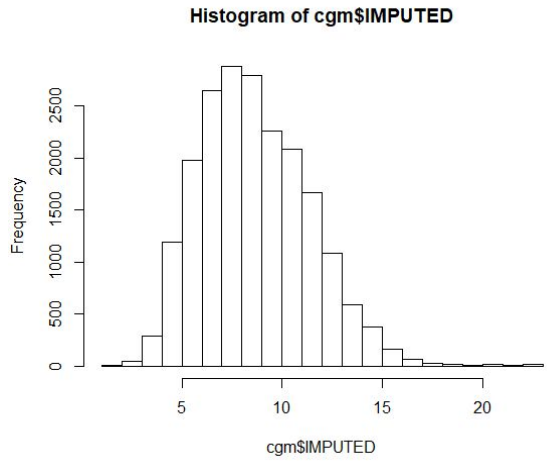
➢ Data stylized facts

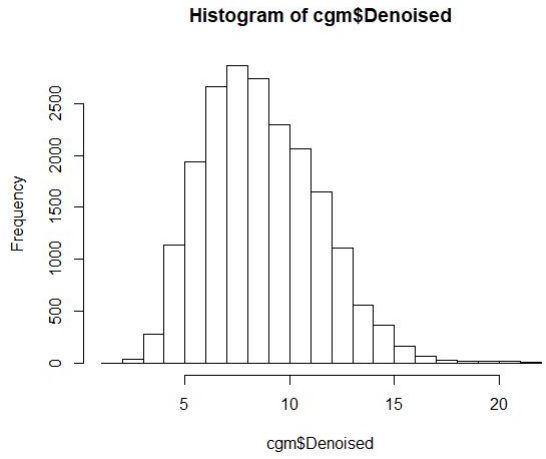
■ Target data (blood glucose values):

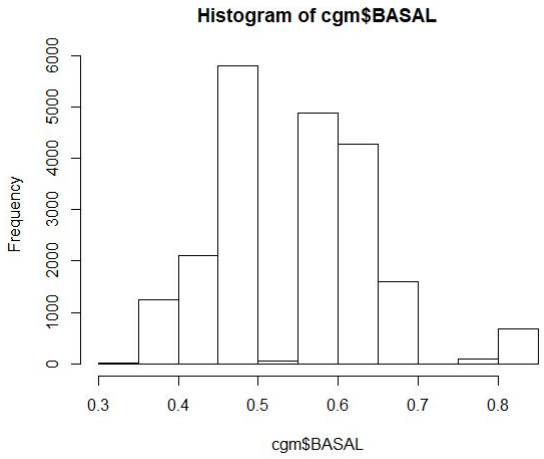


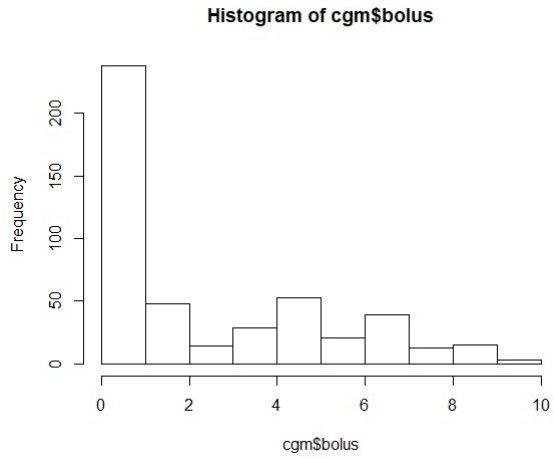


➢ Histograms of some features:









- ➢ Model selection
- ➢ Software packages, applications, libraries, and associated tools, etc
  - R studio: keras, tensorflow, forecast, ggplot2, timetk

❖ Supplemental Project: Pima Indian Diabetes Database

While the project is mostly focused on predicting the tester's blood glucose level by applying univariate/multivariate time series models, we will also be working on Pima Indians Diabetes Database for a classification problem. The goal of this sub-project is to predict whether the patients have diabetes based on the given factors. The model that will be applied to Pima Indians Diabetes Database can be considered as the outcome of the project in case we are not able to produce relatively accurate predictions on Tidepool dataset.

```
-------------------------------------------------------------
Column                        # Observations      Missing
-------------------------------------------------------------
Pregnancies                   768                 0
Glucose                       763                 5
BloodPressure                 733                 35
SkinThickness                 541                 227
Insulin                       394                 374
BMI                           757                 11
DiabetesPedigreeFunction      768                 0
Age                           768                 0
Outcome                       768                 0
```

Variable Name	*Description*
Pregnancies	# of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test

BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin (muU/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DisbetesPredigreeFunction	It provides information about diabetes history in relatives and genetic relationship of those relatives with patients. Higher Pedigree Function means the patient is more likely to have diabetes.
Age	Age (years)
Outcome	Target Variable (0 or 1) where ‘0’ denotes patient is not diabetic and ‘1’ denotes patient is diabetic.