Project One: Diabetes Experiments

Group members:
Yiwen Liu, Michael Lynch, Erzhen Hu, Jinghang Zhang

Part I. Summary of Overall Conclusions

Desiring a way to identify early markers of diabetes, we have modelled the relationships between these substances and their changes over time. We have found that the different features and models perform best for the nighttime fasting setting than for the controlled meal setting, and that different considerations stand out. In the night time setting, it data normalization is important to exclude effects of the differing group means from the data. A subspace discriminant approach worked best here, with 83% classification accuracy against two classes, using a combination of variables that reflect the timing of counteracting processing, the duration of their cycles, and the magnitudes of change at successive intervals. In the setting of a controlled meal, a significant issues was the small size of several of the classes, which required care in dividing the data into testing and training sets. Due to the close similarities between several of the classes, the use of linear discriminant functions, and a visual evaluation thereof, was very helpful and achieved a six-class accuracy of 68% in our tests. However, we believe that making the fine distinctions between all six groups is not critical to the medical needs for early signs of problems. We suggest that a simplification of the problem to three classes, based on glucose tolerance, provides sufficient information for warning signs while allowing for a higher accuracy of 74% using a random forest approach.

Part II. Description of the Problem

The goal of this research is to discover early biomarkers for type 2 diabetes, a very common and debilitating disease. Diagnosis is typically based on high glucose concentrations in the blood while fasting or after a controlled meal. Unfortunately, the standard usage of these tests has little power to identify subjects who show signs of moving towards diabetes despite their normal glucose levels.

The following analysis focuses on the relationships between glucose, insulin, and glucagon. Glucose is a sugar constantly needed by the body for activities. It is obtained directly by eating food, or by creating it in the body from stored fat. When a person eats, the glucose level in the blood spikes. Detecting this elevated level, the body then releases insulin into the blood. Insulin flows through the blood to organ tissue, where it causes the cells to absorb glucose faster for later use. This gradually reduces blood glucose to normal levels. As tissue continues to absorb glucose, concentration in the blood gets low. In response to the low concentration of glucose, the body produces yet another substance, glucagon, which prompts the synthesis of glucose from energy stored in the body. This results in another increase in blood glucose concentration. In a healthy person, the interaction of glucose, insulin, and glucagon is self-regulating. Diabetes is a condition where this process does not properly regulate the glucose level, leading to negative effects of persistently high glucose in the blood.

Our analysis is based on data from two experiments. One study sampled blood from 41 subjects every 10 minutes over the course of 9 hours overnight, while the subject fasted. The other is an experiment in which 76 subjects were given a controlled meal, and had their blood sampled every 5 minutes for 4 hours. In each experiment, due to prior testing, the subjects are known to be diabetic or non-diabetic according to the traditional fasting and glucose tolerance tests. By analyzing the relationships between glucose, insulin, and glucagon in these controlled settings, we hope to identify deeper patterns of interaction that might be early signs of diabetes.

Part III. Descriptions of Methods Not Selected

Methods Attempted for Meal Experiment

For the experiment where subjects were monitored after a controlled meal, we attempted three types of data transformations. First, we used several statistics to describe the whole data set. For the variables, we used the variance and standard deviation of the glucose, glucagon and insulin secretions, the range of each measurement, the time interval between the maximum and the minimum value of each substance secretion measurement, and the maximum and the minimum value of each substance secretion. For these features, we have tried both classification models and linear models. In classification models using all six diagnosis groups (1: NFG/NGT,2:NFG/IGT, 3:IFG/NG,4:IFG/IGT, 5:IFG/DM, 6:DM), we achieved the following accuracy levels: random forest(58.89% accuracy), Boost (52.63% accuracy), KNN (31.58% accuracy), Naive Bayes (42.11% accuracy) and SVM (47.37% accuracy). We draw the conclusion that the Random Forest method works best for the features we extracted from the data set.

The second method we tried is using the first and second derivative of each substance measurement with respect to time. The "Z" secretion values provided to us are a proxy for the first derivative of the concentrations. We then took the difference between two successive secretion measurements at a fixed time interval to get the second derivative. The first derivative and the second derivative contain the information of each measurement's variation with respect to time, and we used them to build the classification model. We tried random forest (59.93% accuracy) and boost (52.75% accuracy). We draw the conclusion that random forest works best with the first and second derivative of measurements.

The third method we used is based on the assumptions of time series. We used time series clustering, wavelet smoothing, and Dynamic Time Warping (DTW) to extract the features and build the classification model. Here we tried both k-fold cross validation (3-fold, 5-fold and 10-fold) on the selected features. However, when we check the histogram of correlations of selected predictors with outcome in **Figure 1 (left panel)** to compare 3 numbers of fold in validation, it turns out that using k-fold cross validation here can result in the missing of at least one class in the data, especially for training data set. In addition, using these transformations, we tried Naive Bayes (42,10% average accuracy, 57.89% maximum accuracy), Random Forest(31.57% average accuracy, 47.37% maximum accuracy), AdaBoost (36.84% average accuracy; 52.63% maximum accuracy), KNN (36.84% average accuracy, 57.89% maximum accuracy),and SVM (36.82% average accuracy, 47.37% maximum accuracy).

Methods Attempted for Nighttime Experiment

For the nighttime fasting experiment, the first features that we used to approximate the variability between substances included the time of the maximum secretion of glucose and insulin, the time of the minimum secretion of glucagon, the lengths of the intervals between these maxima and minima, plus the counts of observations in the top 25% and bottom 25% of the range of each substance. These were our first approximations of the variability and relationships among the substances. A linear support vector machine (SVM) model yielded 76% categorization accuracy.

We developed other features, examples of which are in **Figure 2**, based on identifying upward and downward trending intervals and summarizing their the magnitude, duration, and count of these intervals. Categorization accuracy using these features was quite good, at 93% using fine Gaussian SVM. However, we thought that some of these were too closely related to the underlying differences in glucose mean and standard deviation between the two classes, and so would not be early biomarkers. We attempted to standardize the secretion levels within the diabetic group and the nondiabetic group, and achieved good accuracy (89%). However, we noticed that the subjects misclassified using the raw model were a different group than the ones misclassified using the standardized data. Investigating this led us to conclude that the group-level transformations changed the rank ordering of some of the ratios

that we used as statistics (see **Figure 3** for an example). Consequently, we decided that this group standardization was inappropriate for our features, as it was likely causing problems on edge cases.

Finally, an additional attempt to quantify the relationships between the changing levels of glucagon, insulin, and glucose was to linearly regress the glucose secretion at a given time point onto the glucagon and insulin secretions at the previous time point, and use the regression coefficients as classification features. This did not prove very successful, which we attribute to the poor fits of the curves.

Part IV. Description of the Chosen Methods

Two Approaches to the Meal Experiment

We improved our classification in the meal data using two methods. First, we devised additional features and applied additional algorithms. Second, we combined the groups in a way that simplifies the problem while capturing the essential part of the question.

1. Six-Class Identification

We wanted to improve our six-class categorization accuracy above the levels in the 50-60% range described above. We identified a better approach to segmenting training data and the generation of new features were good opportunities.

With 76 subjects split across 6 classes, several of the classes are fairly small. One class contains only seven subjects. This makes model accuracy highly dependant on how the training and test data are segmented. We had even seen cases where a training set would completely miss a particular class. To ensure representation of each class, we randomly sampled 10% of the subjects from each class and used these as our training set (see **Figure 1, right panel**).

We tried many new transformations of the data. Finally we selected 17 predictors including mean glucose, the standard deviation of three secreta, time to reach the first, second, and third peak of three secreta, time to reach the first, second, and third valley of three secreta, standard deviation of glucose when it reaches its first peak. We plotted the distributions of these features against the six classification groups. The box plot of some selected predictors is shown in **Figure 4.**

We used Linear Discriminant Analysis (LDA) with our chosen features and achieved an average accuracy of 68%, with accuracies as high as 79% based on different sampling seeds. LDA uses Gaussian densities, and the flexible mixtures of Gaussians allow for nonlinear decision boundaries. LDA assumes the classes have a common covariance matrix. It is also applicable in this scenario because of the changing patterns of DM, IFG,and NFG makes some of the classes hard to separate. Additionally, due to the chosen features, some of which contains of peaks and valleys. The LDA output indicates prior probability of groups π=(0.23880597, 0.16417910, 0.08955224, 0.2895522, 0.11940299, 0.17910448). Thus, we can find at most 5 useful discriminant functions to separate the diabetics using the 17 chosen features. **Figure 5** shows the plots of linear discriminants. From these we can see clearly that the LD1 achieves the most separation. This is also evident in the percentages of separation achieved by each discriminant function, in **Figure 6**. **Figure 7** shows a stacked histogram of the values of the first two discriminant function (LD1 and LD2) for the samples from different classes, combined with the scatter plot in **Figure 5**, we can see that LD1 separates sixth group (DM) from other groups well, and it seems to have a cluster for other groups but the boundary is very vague. LD2 separates group 3 (IFG-IGT) and group 2 (NFG-IGT), but it does not separate the other groups very clearly. LD3 separates group1 (NFG-IGT) from group 5 (IFG-DM), and separates group 5 from group 3 but not perfectly. LD4 separates group 5 (IFG-DM) from group 4 (IFG-IGT) but not perfectly. LD5, however, does not separate very well. Therefore, to achieve a good separation of the 6 classes, it is best to use the first four

discriminant functions together. In addition, the proportion of trace shows the proportion of between-class variance explained by these five discriminant functions, with LD1 83.05%, LD2 6.50%, LD3 6.05%, LD4 3.60%, and LD5 0.80%.

2. Merging into Three Groups

In the initial methods for the meal experiment discussed above in part III, the random forest classification model performed better than others, with the confusion matrix reflecting an accuracy of 57.89%. This is not very high; the differences among several of these groups are subtle. We decided to try to reduce number of classes, because the most important value is in finding people on their way towards diabetes, with less importance on identifying their particular class. Because glucose tolerance testing is typically used for evaluation of subjects in meal experiments, we decided to combine groups according to their glucose tolerance level: Normal(NFG/NGT, IFG/NG), Impaired (NFG/IGT, IFG/IGT), and Diabetic (: IFG/DM, DM). We tried the random forest again, and found the accuracy could reach 73.68%. If we assume the probability that people belongs to every group are equal and independent, we can find that the classification of whether a person belongs to a group or not will reach almost 90% which should be a pretty good model.

However, we still have one concern relating to the composition of training data and testing data for groups. Group one (Normal) has more observations than other groups, so it may affect the accuracy of the model. It will be better if we can do the model with more observations and each group can have sufficient people.

Subspace Discriminant in Nighttime Experiment

As noted above, the challenge in the nighttime fasting data is to use features that do not reflect the very wide difference between the means and standard deviations of glucose between diabetics and nondiabetics, and to instead capture some of the self-regulating relationships. To summarize the variations of secretions across the three substances, we started by taking the difference between successive timepoints. This gave us 54 new data points per substance per subject. We defined a "trend interval" as the interval from one local maximum to the next local minimum (or from a minimum to a maximum), counting end points of the series as a maximum or minimum depending on the direction at that point in time. We summarized these by finding the count, average magnitude, and average duration of upward and downward trends per subject per substance. We then made ratios between similar features of different substances, in order to quantify interactions between the substances (see **Figure 2** for examples). In addition to these ratios, we used the differenced data to calculate the mean and standard deviation of differences between successive time points per subject per substance.

After noticing some unanticipated differences in the between-group orderings of ratios calculated from data standardized at a group level (**Figure 3**), we decided to standardize secretion values at the individual level. We subtracted each subject's mean for each substance and divided by the subject's standard deviation for the substance. This centers each individual at mean 0 and casts the secretion levels in units of the subject's own standard deviation. This fully removes the effects of the the glucose means and standard deviations between groups, and leaves measurements that more purely reflect how a given subject's own levels change versus the subject's own mean. We calculated the same statistics on those subject-normalized values.

Using the same three features (median glucagon secretion, ratio of average magnitude of increase in insulin to increase in glucose, ratio of average magnitude of decrease in glucagon to increase in glucose) that performed very well with secretion levels, accuracy topped out at 63% with a linear SVM model. The fine Gaussian SVM models that performed well on the unstandardized data had accuracies of only 44%.

We looked different sets of the features that we had calculated, identifying candidates by examining scatterplots of the distribution of observations in each class. We found the best accuracy by using the time interval between max glucose and max insulin, the ratios of trend durations between insulin and glucose and glucagon and glucose, and the means and standard deviations of the time point to time point differences in glucose and insulin. These gave accuracies as high as 85% for Weighted KNN and 83% for Subspace Discriminant. Given how much raw predictive information we removed by standardization, we think that these accuracy levels are reasonably good, and that the features selected capture most of the important information about the underlying processes of self-regulation. We think that this feature set performed well because it combines measures of the locations in time, information on the relationships between the duration of cycles in the substances, and point to point change magnitudes.

Between the subspace discriminant model and the very slightly more accurate weighted KNN model, we prefer the subspace discriminant model. Having some feature randomization properties in common with a random forest, the subspace discriminant method algorithm selects a random subset of the selected features to train a model, then repeats the process multiple times and gives a final prediction based on the average of the multiple models. As compared to weighted KNN, it had a more favorable false-negative rate (**Figure 8**), and we think that false negatives are more dangerous than false positives in the context of early biomarker detection. There was some (albeit modest) misclassification overlap with the most accurate models on the raw, unstandardized data, which gives us additional confidence that this standardized model is capturing much of the underlying process, without relying on the relatively late biomarkers of the scale differences between groups.

When the model and parameters trained on the night data was applied to the repeat experiment, eight of the ten classifications were consistent between the two versions of the experiment. Its classification of two people changed from experiment to experiment. Unfortunately, when used on the meal data with six categories, these features and this type of model only achieved 34% accuracy.

Part V. Figures and tables

Figure 1: Training and Validation Composition Issues and Methods

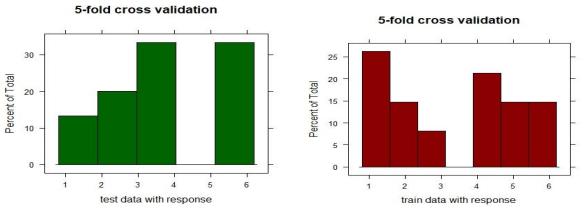
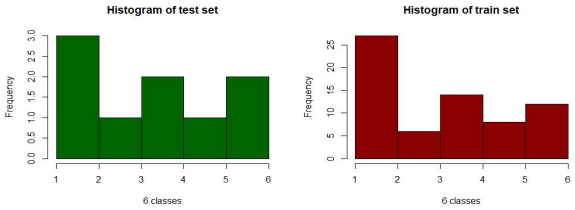
Training and Test Composition: 5-fold CV	**Training and Test Composition: 10% Sample-Based**
	
Using 5 fold validation, some of the test data performs well, but the train data sometimes had a missing class.	The sample based on percentage looks better than the cross-validation ones.

Figure 2: **Examples of Trend and Difference Features in Nighttime Data**

Feature	Definition
ratio_ins_glc_trend_count	number of increasing trend intervals of insulin divided by number of increasing trend intervals of glucose
ratio_ggn_glc_trend_change2	average magnitude of increasing trend intervals of glucagon divided by average magnitude of decreasing trend intervals of glucose
ratio_ggn_ins_trend_duration	average duration of decreasing trend intervals of glucagon divided by average duration of increasing trend intervals of insulin
glc_mean_diff	average of the difference between successive glucose secretions
glc_sd_diff	standard deviation of the difference between successive glucose secretions
Note: A "trend interval" is the interval from one local maximum to the successive local minimum, or vice versa.	

Figure 3: **Magnitude of upward trends in glucose and downward trends in glucagon in nonstandardized and group-standardized data**. Observation numbers are used as data labels. After group standardization, the groups change places, and the cases the observations on the borderline in the nonstandardized data are at the extremes in the group-standardized data.

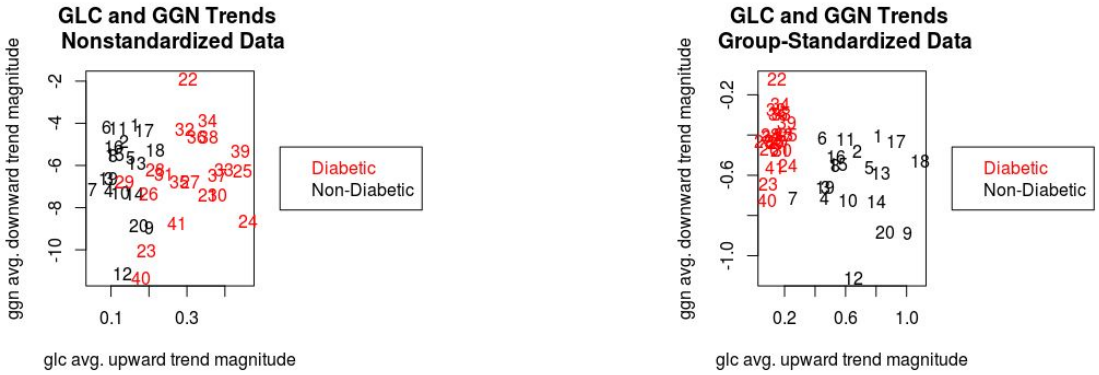
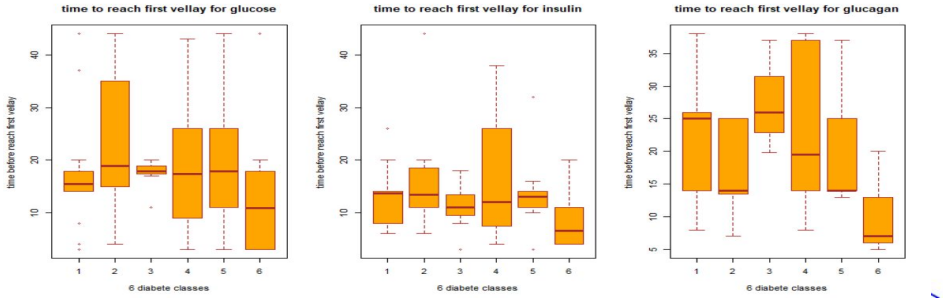


Figure 4



ypredict

	1	2	3	4	5	6
1	2	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	1	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	2

Figure 5: **LDA Plots**

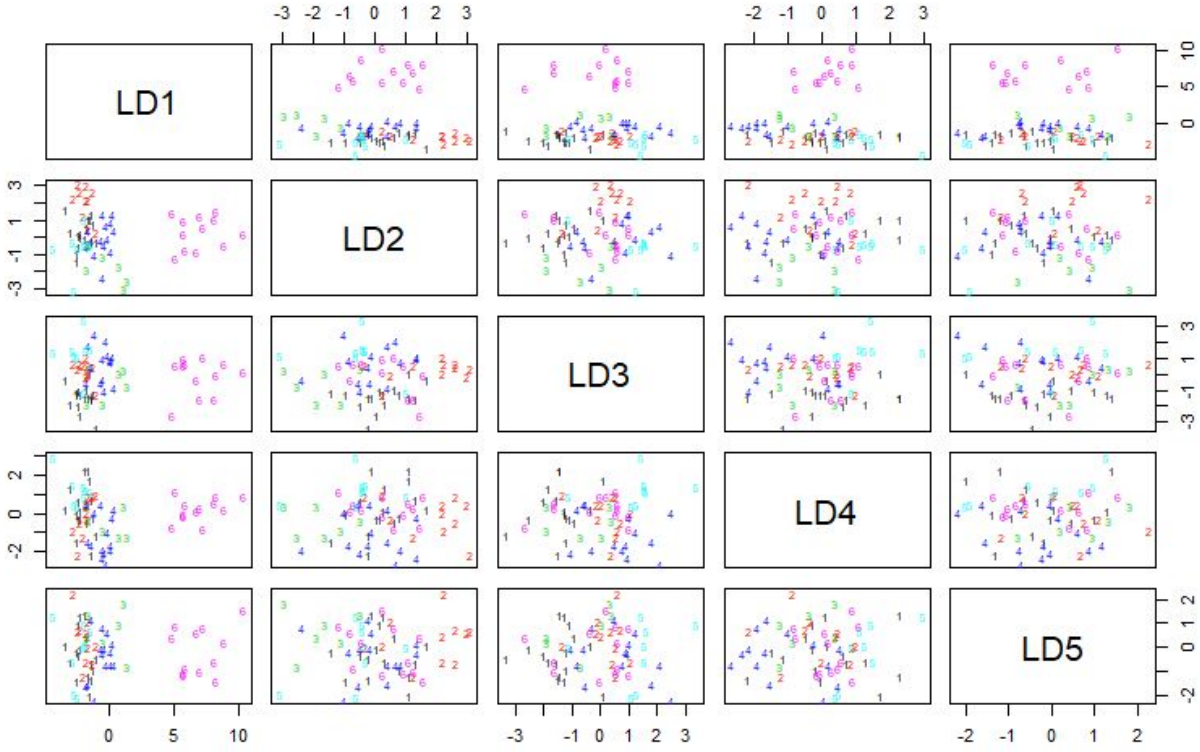
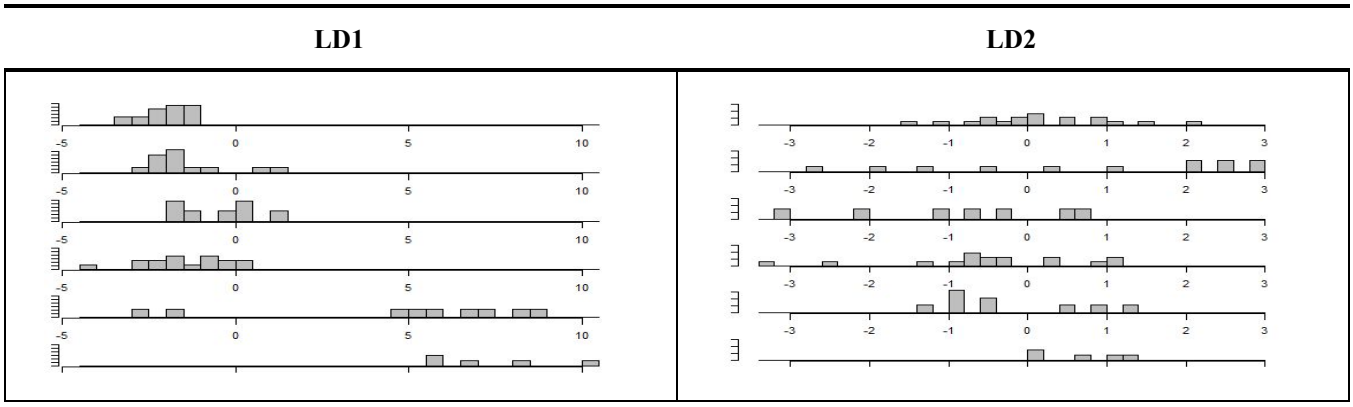


Figure 6: LDA Separation Percentages

LD1	LD2	LD3	LD4	LD5
0.8305	0.0650	0.0605	0.0360	0.0080

Figure 7: Stacked Histograms of LD1 and LD2



LD1 and LD2 performance by stacked histogram(LD1 and LD2 shows most proportion of trace, LD1 83.05%, LD2 6.5%)

Figure 8: Confusion Matrices for Nighttime Data

Subspace Discrimination

		Predicted		
		Nondiabetic	Diabetic	Class Error
Actual	Nondiabetic	16	4	0.2
	Diabetic	3	18	0.143

False Negative Rate: .14

Weighted KNN

		Predicted		
		Nondiabetic	Diabetic	Class Error
Actual	Nondiabetic	20	0	0
	Diabetic	6	15	0.286

False Negative Rate: .29

Contributions of Individual Members

Name	**Areas of Focus & Contribution**
Yiwen Liu	Feature generation, selection, and modelling for meal data. Written contributions to report. Active participation in group meetings.
Michael Lynch	Feature generation, selection, and modeling for night time data. Written contributions to report. Final revisions of document. Active participation in group meetings.
Erzhen Hu	Time series-based and LDA feature generation for the meal data. Written contributions to report. Active participation in group meetings.
Jinghang Zhang	Feature generation, selection, and modelling for meal data. Written contributions to report. Active participation in group meetings.