# Classification of CGM signals based on Machine learning algorithms

Megha Yadav, Projna Paromita and Sahul Phaniraj
*Texas A & M University

***Abstract*—In this work we analyze the performance of various supervised machine learning techniques in classifying the macro nutrient composition of a meal consumed by an individual, based on the individual's Continuous Glucose monitoring (CGM) signals. We conduct a strict binary classification of the macro nutrients in this work. A dataset from PLoS collection has been used for this analysis which comprises of CGM signals from 30 candidates monitored over 3 meals each. Various supervised machine learning models such as Logistic Regression, Support vector Machines, Decision Trees, k-nearest neighbours and Random Forest were used for the binary classification. We also implemented a simple 3-layer Neural network to identify macro nutrient composition of a given meal. SVM performed the best out of all the techniques and provided an overall accuracy of 71 %.**

***Index Terms*—CGM, Bio-feedback system, Supervised learning**

## I. Introduction

Pre-diabetes and diabetes are currently a major health issue. The scale of growth of this disease is global in nature affecting not only individuals but also creating financial impact on national health care systems. [1] [2] American Diabetes Association informs that "The total estimated cost of diagnosed diabetes in 2017 is $327 billion, including $237 billion in direct medical costs and $90 billion in reduced productivity." The alarming rate of increment of economic cost over last 5 years (by 26%) can not be overlooked.[3] Studies on preventing prediabetics to develop diabetes or helping diabetes patients to maintain healthy lifestyle to decrease overall treatment cost and reduced productivity are given priority. Smart technology like continuous glucose monitoring (CGM) devices are used in this process.[4] These are relatively new in the market giving us lots of opportunity to explore the unexplored domain of this signal. The analysis of CGM data allows obtaining more detailed information on glucose variability (GV) patterns during the whole day. In this project, we are particularly interested in monitoring CGM data during the course of the meal.

Healthy diet is one of the cornerstones for preventing diabetes or maintaining a healthy blood glucose level. A prediabetic or diabetic requires to follow some prescribed diet and keep a food journal to track the impact of each meal on their blood glucose level. It is difficult to keep track or know about the macronutrient composition of their meals all the time. Since CGM devices are unobtrusive, cheap and reliable, it would be helpful for patients if there was an application utilizing these CGM devices to keep track of their meal composition on a regular basis. Our objective is to learn the level of macro nutrients in the meal so that the user will have an idea of the macro nutrient composition of the untracked or unknown meal. The first step in achieving this objective would be to provide a model which can classify the macro nutrient composition of a given meal via CGM signals. Therefore, in this work we focus on a binary classification of the macro nutrients present in a given meal.

This paper has been divided into different sections. Section 2 describes about the literature survey. Section 3 deals with Methodology, it further explains the dataset we have used, describes the preprocessing of the data, the features extracted from the dataset and the machine learning models used for classification purpose. Section 4 explains about the results and performs an analysis of the result obtained from different machine learning models. Section 6 and 7 deals with conclusion and references.

## II. Literature Survey

Researchers are focusing on prediction of glucose level using CGM data. Prediction of glycemic level after meal is really important for proper administration of insulin. Incorrect dosage of insulin may result in hypoglycemia or hyperglycemia. This is very crucial for type 1 DM patients and a group of type 2 DM patients who needs external insulin regularly to mainitain the blood glucose level with in the recommended level. There are some specific dosage prescribed by physicians for average people, but it varies from person to person due to many factors related to metabolism. So, personalized prediction of glucose level is given high importance in many research works. In [5], authors have compared 4 nonlinear supervised techniques to predict glucose level for the next 30, 60,90, and 120 minutes after taking meal. Different data assimilation techniques to predict glucose level for type 2 diabetes patients is discussed in [6]. Different genetic programming techniques are utilized to predict glycemia for the the next 30, 60, 90 and 120 minutes after taking meal in [7]. While other papers are working on building personalized prediction models, [8] uses data from other patients to predict glucose level of a person for the next 30 minutes after taking meal using deep network. It also shows how shallow network and deep network prediction varies, resulting in favor of deep network.

Researchers are also focusing on finding hyperglycemic and hypoglycemic periods from CGM data. Prediction of these events is very crucial to prevent possible accidents and major health problems. [9] has shown a way to figuring out abnormal blood glucose level automatically. Moreover, the usage of smart phone devices and deep RNN network to alert abnormal glucose level in case when CGM devices are unavailable to use, is the main focus of this research.

| Nutrients | Bread and peanut butter | Energy bar | Cereal milk and raisins |
|---|---|---|---|
| Calories | 430 | 370 | 280 |
| Fat (g) | 20 | 18 | 2.5 |
| Carbohydrates (g) | 51 | 48 | 54 |
| Protein (g) | 18 | 9 | 11 |

TABLE I: Standardized meal macro nutrient information

As metabolism varies from person to person, personalized therapy and diet is important to pursue healthy lifestyle specially for type 2 DM patients. Most currently available diet charts focus on average population, but it may not satisfy the specific needs of an individual. CGM devices collect enormous amount of data from which we can find pattern related to an individual's diet and lifestyle. We can make changes in an individual's lifestyle and diet and see how it reflects on CGM data to improve their health condition. CGM data along with other parameters like dietary habit, physical activity, gut microbiota is used in [10] to create a personalized diet to maintain blood glucose level within recommended level. [11] introduces Computer Decision Support Systems (CDSS) to help diabetic patients with personalized nutrition and treatment both at hospital and home.

Classifying macro nutrient components of meal using only CGM data is a relatively new field of work, on which we could not find any published research paper. Therein lies the novelty of our work.

## III. Methodology

This section deals with the methodology used in the project. It has been divided into the following sections: a) Dataset, b) Preprocessing, and c) Models

### A. Dataset

The dataset was collected from [12]. Initially they have started with 57 healthy participants without prior diagnosis of diabetes. 30 subjects completed the standardized meal testing portion of the study. Of the 30 subjects, 20 were females and rest are males, aging from 25 to 65. 3 and 7 individuals among them were diagnosed with diabetes and prediabetes, respectively. They have taken 3 standardized meals, each meals twice in the period of the study. These meals vary in proportion of macro nutrients. The glucose concentration was recorded every 5 minutes. On average, 3 hour data, with 30 minutes data prior to the start of each meal until 2.5 hours after the start of each meal data of each subject is used for this work.

### B. Pre-processing and Feature Extraction

CGM is a time-series data. The data set has data points at 5 minutes interval. But some cases, data points were missing. We have interpolated those missing points of the dataset to get a continuous plot. Initially, 177 study days data was available, but among those days, few had less than half hour of data. After discarding those days from our data set, we finally have 171 study days worth data. Then, we have extracted the following 8 features from the data set: 1) Arithmetic Mean, 2) Standard Deviation (SD), 3) Area Under the Curve (AUC), 4) Peak value(Peak), 5) Index of the peak value (MaxIndex), 6) Half-peak Bandwidth (HPBW), 7) Skewness, and 8) last half an hour arithmetic mean (HHmean). The features are calculated as:

$$ArithmeticMean = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

$$StandardDeviation = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}} \tag{2}$$

$$AUC \approx \sum_{k=1}^{N} \frac{f(x_{k-1}) + f(x_k)}{2} \Delta x_k \tag{3}$$

$$Peak = Max(x_i), i \in 1, 2, ...N \tag{4}$$

$$MaxIndex = arg(Max(x_i)), i \in 1, 2, ...N \tag{5}$$

$$HPBW = T_{1^{st}\ 0.5 \times Peak} - T_{2^{nd}\ 0.5 \times Peak} \tag{6}$$

$$Skewness = \gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{7}$$

$$HHmean = \frac{\sum_{i=1}^{N} x_i}{N}, \text{where } i \in \text{ indices in last half an hour time interval of the data} \tag{8}$$

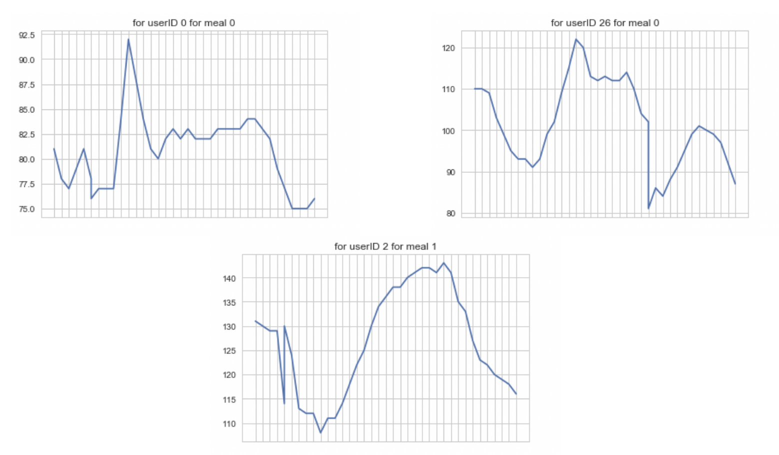

Fig. 1: Various CGM signals of different users for different meals.

We normalize the data (feature column )by scaling between 0 and 1. The normalized value of $e_i$ for variable $E$ in the $i^{th}$ row is calculated as:

$$Normalized(e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

where $E_{min}$ is the minimum value for variable $E$, $E_{max}$ is the maximum value for variable $E$.

We consider three significant macro nutrients viz., Fat, Carbohydrates, and Protein as the output.

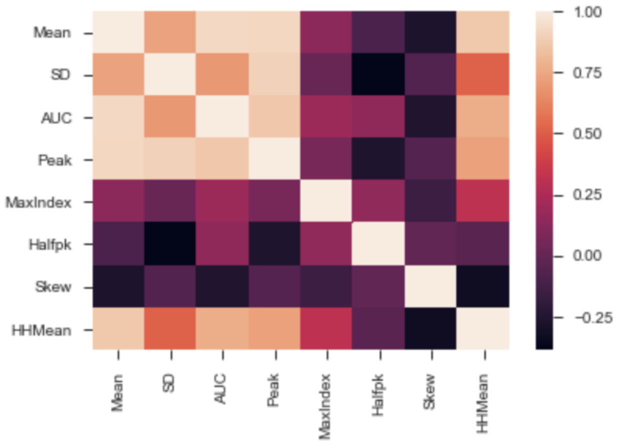

Fig. 2: Correlation Matrix between various classical features

### C. *Models*

We have implemented different models in this project. The classifiers we use are: Neural Network, SVM (linear and RBF kernels), Random Forest, Decision Tree, Logistic Regression and KNN classifier. Since we did not come across any set baseline for the problem statement under consideration therefore, we choose the baseline model as a simple neural network, this network had just 3 layers and 12 nodes in the hidden layer. Neural network gives a good result across all the macro-nutrients.

Feature selection is an integral part of any machine learning model creation, even though we have a limited set of features, we wanted to explore the results with and without feature selection. Therefore, we performed feature selection based on RFE (Recursive Feature Elimination). RFE is performed over linear models like SVM-linear kernel, Decision tree, and Random Forest. In the same context of conducting a through analysis, we tried different variations of the previously mentioned machine learning models by exploring techniques of regularization, boosting etc. For example, for Logistic regression - we created different models which used lasso and ridge regression, also for decision tree - we tried a model with Adaboost and so on.

### D. *Hyper parameter tuning*

Since we are using multiple machine learning models , we have multiple hyperparameters that need to be decided e.g. the depth of a decision tree, Margin tuning parameter for SVM etc. For discovering the appropriate parameters for different ML models, generally cross validation is employed. Since our dataset is user-based, we employed Leave One Subject Out (LOSO) Cross validation. In LOSO CV, we basically consider all user's data except one user's data in the train set and take in the one user's data as the test set. As we can see in the Algorithm 1 , we followed a general psuedocode for the different machine learning techniques, i.e. we started in the outer loop by dividing the train and the test set using LOSO and then in the inner loop for hyper parameter tuning we again utilized LOSO CV to divide the selected train test into train and test set. The ”$some_range$” used for hyperparameter tuning is the various hyperparameter we conduct the LOSO CV over, this range was decided empirically.

**Algorithm 1** Pseudo code for hyper-parameter tuning based on inner CV

```
for $x$ in range $(0, 30)$ do
    //Take one subjects data as test set
    //Take rest 29 subjects data as train set
    //hyperparameter tuning
    for parameter in $(some_range)$ do
        for y in range$(0, 29)$ do
            //Take one subjects data as test set
            //Take 28 subjects data as train set
            train model
            capture accuracy
            update max_acc
        end for
    end for
    append accuracies to array
end for
print average accuracy
```

## IV. Results and Discussion

### A. *Results*

As mentioned previously, we ran 4 parametric machine learning techniques and 2 non-parametric learning techniques. These techniques were used to conduct binary classification of the various macro nutrients present within an individual's meal. We focus on three main macro nutrients in a given meal i.e., Fat, Protein and Carbohydrates. We utilized pre-exisitng libraries from SKLEARN for the implementation of the ML techniques. Since this work focuses on binary classification, the results are based on classifying 3 combinations within each of these 3 macro nutrients i.e low-high, low-med and high-med. This provides us finally with 9 different classification results and accuracies. Therefore, in our final report we average out multiple binary classification results received across one macro nutrient and present these results as seen in Table II.

We considered the neural network results as the baseline. The neural network operated upon raw CGM data and had a simple structure comprising of one input layer, one hidden layer and one output layer. The baseline accuracy thus ranged from $52\% - 54\%$ with the neural network, SVM with a linear kernel provided the best average accuracy of $71\%$ among all the various ML techniques. Logistic regression was the next best with an average $70\%$ overall accuracy. Decision tree and Random forest on an average also provided an average accuracy of $67\%$. The Logistic regression model was constructed with regularization, both ridge and lasso techniques were employed. KNN was slightly better than trees giving an average accuracy of $69\%$.

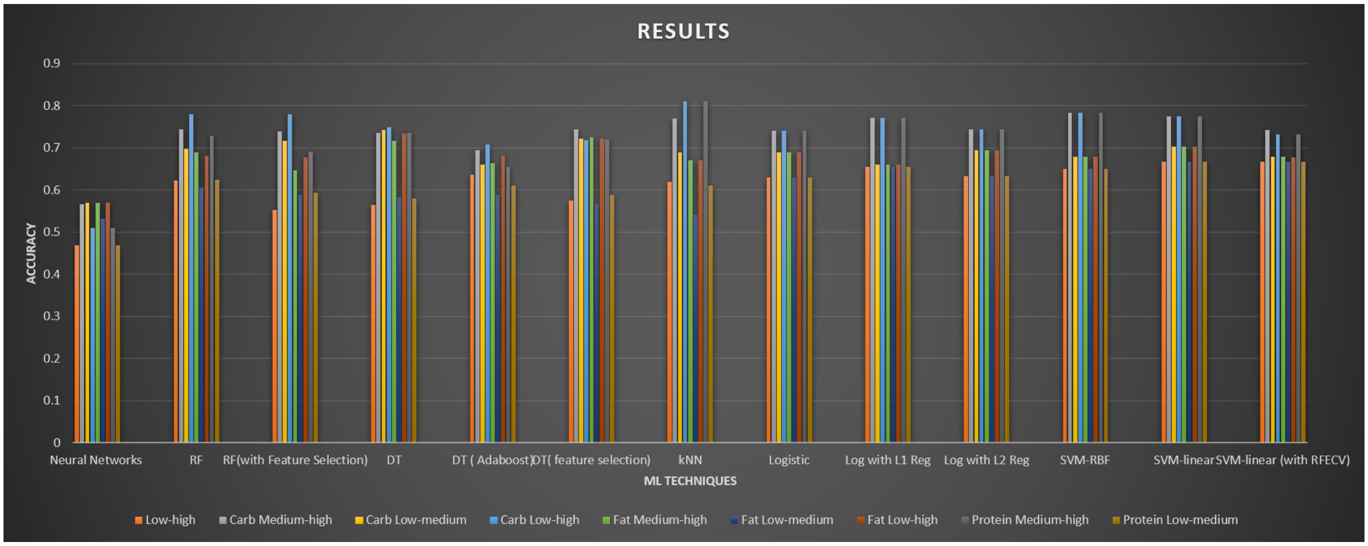

Fig. 3: Results across different Machine learning techniques

| Nutrients | Random Forest | Decision Tree | Neural Network | SVM | KNN | Logisitic Regression |
|---|---|---|---|---|---|---|
| Carb | 0.67 | 0.68 | 0.53 | 0.71 | 0.69 | 0.70 |
| Fat | 0.67 | 0.67 | 0.54 | 0.71 | 0.67 | 0.70 |
| Protein | 0.65 | 0.67 | 0.52 | 0.71 | 0.69 | 0.70 |

TABLE II: Results across 3 meal macro nutrient

### B. Discussion

Except for KNN and SVM RBF kernel the feature selection technique discussed in Models section was applied to all other ML techniques. Feature selection is not supported for KNN and SVM RBF kernel in SKLEARN. This is because KNN and SVM RBF kernel are not model based techniques. One of the challenges of this experiment was a limited data set, we only have on an average 2.5 hours of data per subject, which does not cover the entire effect of the meal consumed. Also, the standardized meals used in the experiment did not vary greatly in their macro nutrient compositions. We also did not have any baseline results to evaluate all the machine learning techniques results . Another challenge faced in this work was the lack of CGM specific features. CGM unlike Speech or other physiological signals does not provide any specific features and therefore we had to resort to utilizing classical time series features, which may not capture CGM specific states.

Since the final result of the classification is one of the three classes of High, Medium or Low therefore it makes sense to implement a tertiary classification instead of a binary classification, this would be taken into consideration in our future works. We also want to explore other ways of extracting features from CGM signals and utilize them for classification. Also, since CGM signals are time series data, we aim to implement a LSTM-RNN model in our future works since it would capture the time based information from a CGM signal.

## V. Conclusion

In this work we conducted a binary classification of various macro nutrients present in an individual's meal based on their CGM signals. For conducting this classification we worked with 8 classical features extracted out of the CGM signals. We tried various machine learning techniques for conducting the classification, out of all the various techniques SVM with a Linear Kernel provided the best average accuracy and f1 score at 71% beating our baseline accuracy provided by the Neural network 54% by a large margin. In future, we aim to conduct a tertiary classification of the macro nutrients investigation with a bigger data set and broader range of features and see how these changes affect the final classification results.

## VI. Team breakdown

The work was in general evenly distributed between all the 3 members of the team. We mention specific details below.
Projna implemented the Decision Tree and Random forest and all variations of these machine learning models.
Sahul implemented the SVM Linear Kernel and SVM RBF Kernel and all variations of these machine learning models.
Megha implemented the Logistic Regression and KNN and all variations of these machine learning models.
Rest all the tasks required to complete this project, presentation and report were divided equally or performed together by all the team members.

## References

[1] L. Guariguata and et al, "Global estimates of diabetes prevalence for 2013 and projections for 2035," ser. Clin. Pract. 103, 137149 (2014), 2014.

[2] J. Beagley and et al, "Global estimates of undiagnosed diabetes in adults," ser. Diabetes Res. Clin. Pract. 103, 150160 (2014), 2014.

[3] [Online]. Available: http://care.diabetesjournals.org/content/41/5/917.long

[4] T. Battelino and et al, "Continuous glucose monitoring in 2010," ser. Int. J. Clin. Pract. Suppl.65, 1015 (2011), 2011.

[5] J. M. Colmenar, S. M. Winkler, G. Kronberger, E. Maqueda, M. Botella, and J. I. Hidalgo, “Predicting glycemia in diabetic patients by evolutionary computation and continuous glucose monitoring,” in *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*, ser. GECCO ’16 Companion. New York, NY, USA: ACM, 2016, pp. 1393–1400. [Online]. Available: http://doi.acm.org/10.1145/2908961.2931734

[6] D. Albers, M. Levine, B. Gluckman, H. Ginsberg, G. Hripcsak, and L. Mamykina, “Personalized glucose forecasting for type 2 diabetics using data assimilation.” ser. PLOS Comput Biol., 2017.

[7] J. M. Velasco, S. Winkler, J. I. Hidalgo, O. Garnica, J. Lanchares, J. M. Colmenar, E. Maqueda, M. Botella, and J.-A. Rubio, “Data-based identification of prediction models for glucose,” in *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO Companion ’15. New York, NY, USA: ACM, 2015, pp. 1327–1334. [Online]. Available: http://doi.acm.org/10.1145/2739482.2768508

[8] H. N. Mhaskar, S. V. Pereverzyev, and M. D. van der Walt, “A deep learning approach to diabetic blood glucose prediction,” vol. 3, 2017, p. 14. [Online]. Available: https://www.frontiersin.org/article/10.3389/fams.2017.00014

[9] W. Gu, Y. Zhou, Z. Zhou, X. Liu, H. Zou, P. Zhang, C. J. Spanos, and L. Zhang, “Sugarmate: Non-intrusive blood glucose monitoring with smartphones,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 54:1–54:27, Sep. 2017. [Online]. Available: http://doi.acm.org/10.1145/3130919

[10] D. Zeevi, T. Korem, and et al, “Personalized nutrition by prediction of glycemic responses,” ser. Cell 163, 10791094, November 19, 2015 Elsevier Inc, 2015.

[11] K. Donsa and et al, “Towards personalization of diabetes therapy using computerized decision support and machine learning: Some open problems and challenges,” ser. Holzinger A., Rcker C., Ziefle M. (eds) Smart Health. Lecture Notes in Computer Science, vol 8700. Springer, Cham, 2015.

[12] H. Hall and et al, “Glucotypes reveal new patterns of glucose dysregulation,” ser. PLoS Biol 16(7): e2005143, 2018.