April 2016

Machine Learning of Lifestyle Data for Diabetes

Yan Luo
The University of Western Ontario

Supervisor
Dr. Charles Ling
The University of Western Ontario

Graduate Program in Computer Science

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Self-Monitoring of Blood Glucose (SMBG) for Type-2 Diabetes (T2D) remains highly challenging for both patients and doctors due to the complexities of diabetic lifestyle data logging and insufficient short-term and personalized recommendations/advice. The recent mobile diabetes management systems have been proved clinically effective to facilitate self-management. However, most such systems have poor usability and are limited in data analytic functionalities. These two challenges are connected and affected by each other. The ease of data recording brings better data for applicable data analytic algorithms. On the other hand, the irrelevant or inaccurate data input will certainly commit errors and noises. The output of data analysis, as potentially valuable patterns or knowledge, could be the incentives for users to contribute more data.

We believe that the incorporation of machine learning technologies in mobile diabetes management could tackle these challenge simultaneously. In this thesis, we propose, build, and evaluate an intelligent mobile diabetes management system, called GlucoGuide for T2D patients. GlucoGuide conveniently aggregates varieties of lifestyle data collected via mobile devices, analyzes the data with machine learning models, and outputs recommendations.

The most complicated part of SMBG is diet management. GlucoGuide aims to address this crucial issue using classification models and camera-based automatic data logging. The proposed model classifies each food item into three recommendation classes using its nutrient and textual features. Empirical studies show that the food classification task is effective.

A lifestyle-data-driven recommendations framework in GlucoGuide can output short-term and personalized recommendations of lifestyle changes to help patients stabilize their blood glucose level. To evaluate performance and clinical effectiveness of this framework, we conduct a three-month clinical trial on human subjects, in collaboration with Dr. Petrella (MD). Due to the high cost and complexity of trials on humans, a small but representative subject group is involved. Two standard laboratory blood tests for diabetes are used before and after the trial. The results are quite remarkable. Generally speaking, GlucoGuide amounted to turning an early diabetic patient to be pre-diabetic, and pre-diabetic to non-diabetic, in only 3-months, depending on their before-trial diabetic conditions. cThis clinical dataset has also been expanded and enhanced to generate scientifically controlled artificial datasets. Such datasets can be used for varieties of machine learning empirical studies, as our on-going and future research works.

GlucoGuide now is a university spin-off, allowing us to collect a large scale of practical diabetic lifestyle data and make potential impact on diabetes treatment and management.

Keywords: machine learning, lifestyle data, T2D management, mobile computing

Acknowlegements

Foremost, I would like to thank my supervisor sincerely, Prof. Charles X Ling, for his continuous support, guidance, and insightful advice. Next, I want to thank Dr. Petrella and Jody Schuurman for organizing the clinical trials. Last but not least, my acknowledgement and love also go to my parents, Yang Luo and Huiyan Zhu, who have brought me to this wonderful world, taught me to appreciate everything I have, directed me through ups and downs and stood by me at all time.

My research is supported by NSERC Grants and scholarship from the University's the School of Graduate Studies. This thesis would not have been possible without the generous resources provided by the Department of Computer Science.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

Diabetes is a metabolic disease in which patients have abnormally high blood glucose. There are two main types of diabetes: Type-1 and Type-2 Diabetes. Type-1 Diabetes (T1D) is a disease in which the pancreas produces little or no insulin. As such, glucose builds up in blood instead of being used for energy. Individuals with T1D need to inject insulin as prescribed. Type-2 Diabetes (T2D), on the other hand, is a disease in which the pancreas does not produce enough insulin, and/or the human body does not properly use the insulin it makes [2, 3]. There is also a type of diabetes called Prediabetes, which refers to blood glucose levels that are higher than normal, but not yet high enough to be diagnosed as T2D [2].

Diabetes, if not well-treated, will develop many long-term complications including heart attack, stroke, amputations, and blindness. According to WHO 2013, 347 million people worldwide have diabetes, and T2D patients comprise more than 90% of them. Thus, improving the treatments of T2D is significant to patients' condition and our society [2].

After a person is diagnosed with T2D, usually medications will be prescribed by doctors. Also, a protocol called Self-Monitoring of Blood Glucose (SMBG) will be advised by healthcare providers to help the patient achieve proper blood glucose targets [2]. In fact, blood glucose is affected by a large number of lifestyle and clinical factors, including what you ate, how long ago you ate, your previously blood glucose levels, physical activity, mental stress, illness, sleep patterns, and so on [2]. Each factor can have different impact on a specific person's blood glucose and there could also be complex interactions among these factors.

1.1 Challenges and Barriers of Diabetes Management

Adopting and maintaining healthy lifestyle changes is highly challenging for T2D patients [4]. For example, the overwhelming complexity of carbohydrate or calories counting presents an often insurmountable obstacle for most T2D patients. Also, when the patient's blood glucose is high, it can be difficult to determine which factor(s) causes it. On the other hand, the time-constraints for healthcare providers do not allow for 24/7 real-time monitoring and personalized advice, leaving a patient in a potentially life-threatening situation.

The rapid development of mobile computing starts a new era of diabetes management. Mobile apps and wearable medical devices have significantly facilitated diabetes management regarding their usabilities. For example, it is possible now that patients can conveniently sync their blood glucose readings from Bluetooth-enabled glucometers to their mobile apps and send them to health providers via the Internet. Their health providers can log-in Web portal to review and analyze their blood glucose readings. Such ubiquitous procedure is a revolution. However, there are still two main challenges preventing the current mobile diabetes management systems from reaching their full potential.

The first challenge is related to the system input, i.e., the complexities of lifestyle data logging. Even utilizing the modern mobile diabetes system and wearable sensors, data logging is still troublesome and time-consuming, especially for diet and glucose levels. For example, most systems use food database searching to help patients record diet information. However, to record every detail of a meal, patients need to search each food item in the meal and estimate its serving and portion size, which makes diet recording a very time-consuming process. As for blood glucose levels, continuous blood glucose monitoring (CGM) devices or finger-prick based glucometers are mostly used to estimate glucose concentration from real blood samples, which means testing blood glucose level is painful and costly [20, 59].

The second challenge is related to the system output. More specifically, most systems are lacking evidence-based, personalized, and instantaneous feedback to patients' diabetes management practices. Diabetic lifestyle data now are collected without being adequately analyzed. Some popular systems such as Glooko ™ and Glucose Buddy ™, only log and plot blood glucose and other lifestyle data, without data analytics on patient data to actively advise and engage patients. A few others, such as WellDoc [5], do give patients personalized advice but only prescribed by doctors who need to access the system to review their patients data. That is in fact not practical when the data volume is huge and data structure becomes complex.

1.2 Our Solutions

A trend in healthcare analytics is taking advantage of machine learning models to discover patterns from the large scale of clinical data and to aggregate medical domain knowledge from a variety of sources. Specifically, predictive modeling, as an important component of machine learning, is a process used in predictive analytics to create one or multiple models with forecasting probabilities and trends. It could be utilized to facilitate the processing of patients' data and allow patients and their health providers to interact in a more convenient and timely manner.

To tackle the data input issue. We believe that machine learning technologies can be utilized to facilitate lifestyle data input. For example, it is possible now that patients can take pictures of a food item, and that food item can be recognized using predictive models such as deep neutral networks. Then, we can use the food meta-data learned from machine learning models to obtain detailed nutrition information from querying food databases.

Machine learning could also be used to enhance the outputs of diabetes management system, i.e., learning of lifestyle data and output practical hidden patterns or knowledge. Numerous research works have been proposed to use predictive models in diabetes diagnosis and risk analysis [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. They aim to predict diabetes clinical diagnosis and risk factors based on patients' profile data. Another challenging research area is blood glucose prediction. With the help of reliable blood glucose prediction models, patients would be able to take actions proactively to stabilize their predicted blood glucose levels. Previous studies have been focusing on the exclusive CGM (Continuous Blood Glucose Monitoring) data, and achieve remarkable prediction accuracy [16, 17, 18, 19]. To elaborate, CGM is an advanced way to measure glucose levels in real-time (or short sampling interval like 5 minutes) throughout the day and night. In particular, glucose sensors are inserted under the skin to measure glucose levels in tissue fluid. They are also connected to a transmitter that sends the information via wireless networks. Such procedure mostly used in clinical treatments for T1D patients, which are usually costly and publicly inaccessible.

As for T2D patients, we believe that the design principle of blood glucose prediction framework should focus on the lifestyle data and discrete finger-prick blood glucose samples since they are the essences of T2D self-management. Furthermore, the lifestyle data nowadays can be collected cost-efficiently (comparing to the clinical data such as CGM data) using modern mobile devices. The lower of the cost of lifestyle data collection makes it has the potential of being expanded to big informative data.

In this thesis work, we propose, build, and evaluate an intelligent mobile diabetes management system, called GlucoGuide for T2D patients. GlucoGuide conveniently aggregates varieties of lifestyle data collected via mobile devices/sensors, analyzes the data with predictive models, and outputs the model outcomes as recommendations.

The most complicated part of self-management is diet management, and GlucoGuide aims to address this crucial issue using classification models and mobile-camera-based automatic data logging. The proposed classification model can categorize any food item into three recommendation classes using its nutrient and textual features. Our evaluation shows that the best model classification accuracy is around 95%, indicating this method is empirically effective and reliable for food items recommendations.

A lifestyle-data-driven recommendations framework in GlucoGuide can output short-term and personalized recommendations for lifestyle changes to help patients stabilize their blood glucose level. This framework conduct predictive modeling on patients' lifestyle data to predict blood glucose levels. Each prediction iteration can also rank lifestyle factors in terms of their importance or relevance. The data-driven recommendations regarding the important factor(s) will be sent to patients.

To collect training data from real patients and evaluate the clinical effectiveness of this framework, we have collaborated with Dr. Petrella's health team to conduct a three-month clinical trial on human subjects. Due to the high cost and complexity of trials on human, a small but representative subject group is involved. Using the lifestyle data collected from this group, the performance of our blood glucose prediction framework is similar to state-of-the-art methods [20, 16] using clinical CGM data.

In terms of clinical effectiveness, two standard laboratory blood tests for diabetes are used before and after the trial. The results are quite remarkable. Generally speaking, GlucoGuide amounted to turning an early diabetic patient to be pre-diabetic, and pre-diabetic to non-diabetic, in only 3-months, depending on their diabetic conditions before the trial.

Since the collected clinical dataset is relatively "small“ regarding its volume and dimensions, our blood glucose prediction framework was possibly not converged. Meanwhile, we do not have large scale of practical diabetic lifestyle datasets to explore the potential of our model at this point. As such, we have designed an effective mechanism to expand and enhance this clinical trial dataset. After the data expansion, we have obtained large scientifically controlled artificial datasets for further studies on blood glucose predictions.

Specifically, we first use parametric statistical inference to find the properties of underlying distributions and potential dependencies. We have also applied diabetic domain knowledge to make certain assumptions on the data and determine all the well-known feature dependencies and parameters. The unlabeled lifestyle datasets will be sent to AIDA (a very popular online diabetes simulator) to generate blood glucose labels [75, 76]. Our evaluation results show that this generation process is effective. Such artificial datasets can also be used for varieties of machine learning empirical studies such as blood glucose warnings, missing value imputation, data sparsity analysis, transfer learning, etc., as our on-going and future research works.

GlucoGuide now is a university spin-off, allowing us to collect a large scale of practical diabetic lifestyle data and make potential impact on diabetes treatment and management. As far as we know, diabetic datasets similar to ours are not publicly available at this point, thus, we also plan to release our de-identified datasets for better collaborations with peer researchers.

The remaining chapters of the thesis are presented as follows: We first review the prior relevant research and developments on mobile diabetes management system and diabetes modeling using predictive models in Chapter 2. Then, our research question is defined in Chapter 3, as well as the design principle of GlucoGuide system. The diet management sub-system and food classification of GlucoGuide is explained in Chapter 4. Chapter 5 presents the proposed diabetic lifestyle analytical framework and its empirical and clinical evaluation results. The clinical data expansion and enhancement process is presented in Chapter 6.

The current research and development status of GlucoGuide spin-off and its future research objectives are briefly discussed in Chapter 7. The thesis is concluded in Chapter 8.

Chapter 2

Literature Review

In this chapter, we will first discuss the features of diabetic lifestyle data and how they can be collected and managed via mobile diabetes management systems. Next, we will review relevant machine learning researches on diabetes-related predictive tasks such as diagnosis and risk analysis.

2.1 Mobile Diabetes Management System

Nowadays, mobile computing has been emerging as an integral part of daily activities. As for their application in medical informatics, the technological advances could make medical services and treatment overcome time and location barriers thus provide ubiquitous, real-time, individualized medical treatments [21, 22, 23]. In fact, researchers have consistently shown that such ubiquitous healthcare services could be utilized to enhance the quality of life for people living with chronic illnesses (such as diabetes, heart disease, cancer, and so on.) [24, 22].

It is well known that the healthcare for T2D is mostly focused on lifestyle changes and self-management, which are typically managed by increasing exercise, diet modification, and/or medications/insulins adjustments. If blood glucose levels are adequately lower than certain levels, medications such as metformin or insulin might not be needed.

In fact, self-management is often a challenge for T2D patients since the control is contingent on numerous complex factors and behaviors. One promising solution is to integrate ubiquitous healthcare services into diabetes self-management, which could bring more effective behavioral changes and continuous health monitoring [25].

2.1.1 Features of Diabetes Lifestyle Data

Health features are aspects of the health conditions that need to be managed and monitored to best control the chronic illness [26]. Smith and Schatz [26] identify 14 health features, which are Access to care, Air quality, Blood pressure, Blood glucose levels, Depression, Diet, Education, Heart rate, Medication, Physical Activity, Respiration, Stress, and Substance abuse respectively. In terms of diabetes, the most relevant health features of lifestyle data are listed as follows (other similar health features for diabetes can also be found in [24, 27, 5, 28]):

- *Blood Glucose*. Blood glucose level is the most important feature and the target variable for diabetes [25]. This feature is typically collected via either finger-prick glucometer or Continuous Glucose Monitor (CGM) [26, 29].The blood glucose readings can be transmitted to mobile devices wirelessly or using USB port (some advanced glucometers follow Bluetooth communication specification made by Continua Health Alliance). To give an example, Cho et al. [29] employed mobile phone to collect blood glucose measurements. Their results indicated good levels of participants' satisfaction and adherence using the mobile technologies to manage blood glucose levels. More conveniently, in the future, we would image the appearance of wearable sensors measuring glucose levels without actually taking blood glucose sample, such as Google's prototype called "Smart Contact Lens" for measuring blood glucose levels.

- *Diet*. Diabetic patients need to delicately monitor carbohydrates intake to avoid abnormal high/low blood glucose level or adjust insulin dosage (mostly for T1D patients). Other nutrients information such as protein, fat, sodium could also affect their blood glucose level. Researches have also shown that mobile nutritional support is effective on diabetes management [26, 29]. Nowadays, diabetes patients could conveniently use advanced technologies such as barcode scan, Optical Character Recognition (OCR), or food database searching to conduct diet management. Shortly, with the rapid developments of advanced machine learning technologies, especially for deep learning, meals could be recorded via image or voice recognition reliably.

- *Physical Activity*. Physical activity is typically measured by exercise intensity and duration [25]. Step counts tracking (pedometer) is another convenient method for estimating the exercise calories consumption. It is proved that moderate or vigorous exercises could increase the insulin sensitivity of cells. Thus, keeping track of exercise is important to T2D patients. Stuckey [30, 31] et al. conducted The Diabetes and Technology for Increased Activity (DaTA) study to test the effectiveness of a lifestyle intervention driven by self-monitoring of blood glucose (BG), blood pressure (BP), physical activity (PA), and weight using the mobile devices and wireless communication technologies. Their research results showed that self-monitoring of the risk factors for metabolic syndrome (such as hypertension and dysglycemia) and increased physical activity improved the participant's cardiovascular disease risk profile. Furthermore, with the help of hourly step counts information and aerobic step counts, future systems can make more specific and intelligent recommendations to improve patient's self-management.

- *Access to Care.* Access to care mainly refers to the capabilities of being able to communicate with the health providers in the case of questions or emergency. This can be done via voice, messages, E-mails between mobile devices and servers [29, 30, 22, 32, 33, 34]. This feature is important for all chronic self-management systems. Patients are motivated to upload their data since they know that all data could be carefully taken care by health providers, and they would receive valuable feedback. The location-awareness mobile device can also report patient's location to health provider to deal with the emergent situations.

- *Blood Pressure.* Blood pressure can be measured using wireless communication enabled sphygmomanometer or other wearable devices such as Fitbit.

- *Weight.* Weight is normally related to BMI measurement. Bluetooth-enabled scales can conveniently record weight data and communicate with mobile phones.

- *Education.* Education is important for diabetes patients since self-management involves many complex factors and most diabetic patients lack such knowledge. Mobile devices have been proven very useful in this area [26]. However, according to the best of our knowledge, personalized and real-time education is still a big challenge and largely unexplored.

- *Substance abuse.* Substance abuse normally refers to tobacco and alcohol. Mobile devices could use message, reminder, or alert functionality to encourage smoking cessation and reduction in alcohol.

Current mobile diabetes systems are designed to facilitate the management of these health features. On their clinical effectiveness, Liang et al. [35] reviewed 22 clinical trials that assessed the effect of mobile phone intervention on blood glucose control of patients with diabetes. Most results show that significant reduction in glycosylated hemoglobin $A1_c$ values and / or other health outcomes in diabetes patients when comparing the mobile phone intervention group with the control group. We list several landmark research papers and clinical trials which represent the typical research methodologies and evaluation procedure of ubiquitous diabetes management research:

- Tang et al. [27] evaluated an online disease management system supporting patients with uncontrolled T2D. It is a randomized controlled trial of 415 patients with T2D. Patients in the intervention group (INT) need to upload their glucometer, nutrition, exercise logs, and insulin record. Nurses and physicians remotely provide glucose summary report, personalized text and video, and diet advice to INT group. Its clinical trial results show that INT group has significantly reduced $A1_c$ at six months compared to usual care (UC) (-1.32% vs. -0.66%). While for the 12 months duration, the difference was not very significant (-1.14% INT vs. -0.95% UC). According to their analysis, the main potential reason is that UC group could have stimulated behavioral changes due to “Hawthorne effect“ (changes influenced by being observed in a study).

- Quinn et al. [5] conducted a similar 3-month study. The INT group received mobile phone based software which provides real-time feedback on patients’ blood glucose level, medication regiments, incorporated hypo- and hyperglycemia treatment algorithm and other requested additional feedback. The system sent computer-generated logbooks with suggested treatment plans to INT group. They claim that adults with T2D achieved statistically significant improvements in $A1_c$ as well as satisfaction levels.

- Noh et al. [36] designed and developed a web-based ubiquitous information system for diabetes education. They compared the INT group using this system with conventional education for diabetes patients. Their results show that blood glucose level and $A1_c$ are significantly decreased over time in the INT group but not in the control group.

- Spring et al. [37] argued that some patients exhibit multiple chronic disease risk behaviors. They believe that the current literature provides little information about advice that can maximize simultaneous health behavior changes. They randomized 204 adults with observed unhealthy behaviors such as saturated fat and low fruit and vegetable, high sedentary leisure time, and low physical activity. They were requested to use mobile devices to record and upload their daily activities. The increase fruits/vegetables and decrease sedentary leisure treatments improved more than the other three type of treatments ($P \leq 0.001$). Specifically, daily fruit/vegetable intake increased from 1.2 servings to 5.5 servings, sedentary leisure decreased from 219.2 minutes to 89.3 minutes and saturated fat decreased from 12.0% to 9.5% of calories consumed.

- In the pilot-controlled clinical trial conducted by Katz and Nordwall [34], they utilized Bluetooth, mobile phones, and application servers to facilitate the self-management pro-

> cess. More importantly, their system translated scientifically supported knowledge for chronic disease management into action by providing easily followed daily coaching using the patients' data. It is a data-driven system because the feedback messages are selected by the system based on the patient's historical data. However, their feedback is simply generated using domain knowledge and needed human involvement to review the data and provide personalized feedbacks. We argue that such process is not feasible when the data scale become large.

We have surveyed state-of-the-art products and applications in the industry. By reviewing their product specifications, we can conclude that all the *health features* mentioned above are manageable via nowadays' mobile technologies and Web technologies. Most recently, IT giants such as Google, Apple, and Microsoft have also released their mobile health platforms for fitness and chronic diseases managements. Take ResearchKit for an example, which is an open source framework introduced by Apple to enable efficient lifestyle and clinical data collection and aggregation. Researchers can easily create visual consent flows, conduct real-time dynamic active tasks and survey using a variety of customizable modules provided by ResearchKit. Also, since ResearchKit works seamlessly with HealthKit, allowing iOS apps that provide health and fitness services to share their data with each others. Researchers thus can access even more relevant lifestyle data for their studies, such as daily step counts, calorie use, and heart rate, etc.

Although the clinical trials and products mentioned above show that it is technically feasible to improve diabetes self-management using mobile devices and Web technologies. However, most systems are very similar in their functionalities and implementations, and there are obvious gaps between the evidence-based recommendations and their current implemented functionalities. Furthermore, in spite of the effectiveness of these clinical trials, most of them are not cost-effective. Participants in the clinical trials still take large efforts to synchronize their lifestyle data. Health providers on the server-side need to routinely check and review patients' data and provide feedback. As such, we believe it is still very challenging to design and develop an intuitive, intelligent, and interactively ubiquitous diabetes management system.

Predictive models, as an important component of machine learning, could be utilized to facilitate the data processing and allow patients and their health providers interact in a more convenient and timely manner. Researchers have been exploring the application of predictive models in diabetes domain for nearly two decades. In the next section, we will discuss how predictive modeling can be applied to diabetes diagnosis and risk analysis.

2.2 Diabetes Risk and Diagnosis Using Predictive Modeling

Predictive modeling, also can be called supervised learning, is a process used in predictive analytics to create one or multiple models with forecasting probabilities and trends. It has two main components: Classification and Regression. Classification aims to categorize data entries into discrete class labels (e.g., Spam vs.Non-Spam), with minimum classification errors subject to applicable constraints. Regression tries to predict continuous values such as stock price, temperature, etc. The overall performance of regression is usually measured by mean absolute error or mean squared error.

Diabetes risk and diagnosis prediction are typically considered as classification tasks, classic classification framework such as Support Vector Machine (SVM) [6, 10, 15], Decision Tree [12], RandomForest [14], Neural Networks [9], Naïve Bayes [11] are widely employed. Training datasets are focused on clinical diagnoses and patient's profile data, including patients' family history, age, race and ethnicity, weight, height, waist circumference, body mass index (BMI), hypertension, diagnosis code, etc. Once classification models are well-built, it is clearly to predict and identify the important rules and features causing diabetes disease.

Yu et al. [10] built SVM models on patient data from the 1999-2004 National Health and Nutrition Examination Survey (NHANES). They designed two classification schemes: Classification Scheme I (diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes) and Classification Scheme II (undiagnosed diabetes or prediabetes vs. no diabetes), and achieved receiver operating characteristic (ROC) curve, were 83.5% and 73.2%, respectively. Nahla et al. [15] used an additional explanation module, turning the “black box“ model of an SVM into an intelligible representation (ruleset). Evaluation results on a real-life diabetes dataset show that their intelligible SVMs achieve with the prediction accuracy of 94%. Polat et al. [6] have used Generalized Discriminant Analysis to discriminant features between healthy individuals and patients (diabetes). Then, they built LS-SVM to classify diabetes dataset. The proposed system called GDALS-SVM and obtained 82.05% classification accuracy using 10-fold cross validation. Similar research classification works have also been found in the literature using Decision Tree [12], Artificial Neural Network (ANN) [9], and Naïve Bayes [11].

Li and Zhou in [14] argued that if we can learn models in the presence of a large amount of undiagnosed samples, the heavy labeling burden on the medical experts could be released. They proposed a semi-supervised learning algorithm named Co-Forest. It extends the co-training paradigm by using a well-known ensemble method named Random Forest, which enables Co-Forest to estimate the labeling confidence of undiagnosed samples and produce the final hypothesis easily. Their diabetes case studies show that undiagnosed samples were helpful in building computer-aided diagnosis systems, and their model was able to improve the performance via learning only a small amount of diagnosed samples by utilizing the available undiagnosed samples.

Although the diabetes diagnosis and risk analysis are well-studied using predictive models, we argue that more short-term and actionable predictive tasks could be performed. For example, predictive models could be used to learn the knowledge from dietitians and then provide food guide for T2D patients. More advanced, given sufficient training data, predictive models should be able to discover which lifestyle factor has the largest influence on the abnormal blood glucose level.

In the next chapter, we will further discuss our research questions regarding the challenges and barriers of current diabetes self-management and how predictive models can be applied to solve them.

Chapter 3

Overview of the GlucoGuide System

In this chapter, we will further discuss the main challenges of current mobile diabetes management systems. Next, we will compose our research questions regarding the solutions to these challenges via applying machine learning technologies and mobile computing.

3.1 Challenges

Based on the remarks of literature review and our preliminary studies, we think there are two main challenges for current mobile diabetes management systems. The first one is the complexities of lifestyle data input. The second one is the lacking of evidence-based lifestyle recommendation for patients to effectively stabilize their blood glucose. These two issues are in fact connected and affected by each other. The ease of data recording enables more and better quality of training data from patients for applicable predictive models to generate better results. On the other hand, the irrelevant or inaccurate data input will certainly commit errors and noise to further data analysis. The output of predictive models, as potential personalized patterns or knowledge, could be incentives for users to record more qualified data.

Most lifestyle data nowadays can be automatically captured via a variety of sensors (such as glucometers, pedometers, blood pressure monitors, etc.). Smart wearable sensors like Fitbit products can also collect comprehensive health data about your body, especially for your physical activities. However, for diabetes patients, logging diet and glucose level are still relatively difficult. As discussed before, machine learning technologies such as most recent deep learning can be utilized to help patients with diet and blood glucose input. For example, it is possible now that patients can take pictures of a food item, and that food item can be recognized using deep neutral networks. Then, we can query the meta-data about that food item in a food database to get its nutrition information.

How to tackle the second challenge is the focus of this thesis work. We aim to collect and analyze lifestyle data using predictive models. Based on the model outcomes, evidence-based, personalized, and timely recommendations could be generated to guide patients' self-management practices. Such recommendations are closely related to patients' data, thus can effectively evaluate their performance and help patients creating more specific and personalized action plans. On the other hand, recommendations can also be viewed as incentives of data recording and motivate patients to record and upload more qualified data.

3.2 Research Questions

We set the focus of this thesis on T2D patients since it comprises more than 90% of the diabetic population [25]. Thus, improving the treatments of T2D is significant to patients' condition and our society. We believe predictive models could be applied to the two most important and complex components of T2D self-management: 1) Diet management and 2) Blood glucose monitoring and predictions. Specifically, predictive models could be used to analyze food data and categorize each food item into different recommendation levels for patients. More intelligently, predictive models could also be used to predict blood glucose level and identify what factor(s) causes abnormal blood glucose levels. Since T2D patients usually do not use CGM devices, our blood glucose prediction task will focus on lifestyle data and fingerstick-based blood glucose samples.

As such, we propose three research questions or hypotheses as follows:

- **Can we use classification models to generate real-time food guideline to help patients proactively manage their diet?**

- **Can we predict T2D patients' blood glucose level merely based on lifestyle data and discrete fingerstick-based blood glucose samples?**

- **Can we provide clinically effective lifestyle recommendations based on the outcomes of blood glucose prediction?**

To answer these research questions, we propose, build, and evaluate an intelligent mobile diabetes management system, called GlucoGuide for T2D patients. GlucoGuide conveniently aggregates varieties of lifestyle data collected via mobile devices, analyzes the data with predictive models, and outputs recommendations.

In the next section, we will present the architecture of GlucoGuide.

3.3 The GlucoGuide Solution

GlucoGuide contains three sub-systems, as shown in Figure 3.1, including the 1) Lifestyle data collection and preprocessing, 2) GlucoGuide mobile clients, and 3) GlucoGuide Engine. Their design and development processes involve many up-to-date technologies.

The GlucoGuide mobile clients has been designed and built on the latest Google Android and Apple iOS systems, aiming to provide excellent user experience powered by our well-designed user interface and information visualization technologies. One design goal here is to reduce the burden of lifestyle data input in GlucoGuide. For example, we have been designing and implementing the camera-based auto logging of food items and glucometer readings. It acts as a bridge connecting patients and GlucoGuide Engine. The mobile clients collect raw lifestyle data, preprocess and upload them securely to GlucoGuide Engine where data analysis takes place. Patients can also use the mobile clients to receive and review recommendations, setup different kinds of reminders, etc.

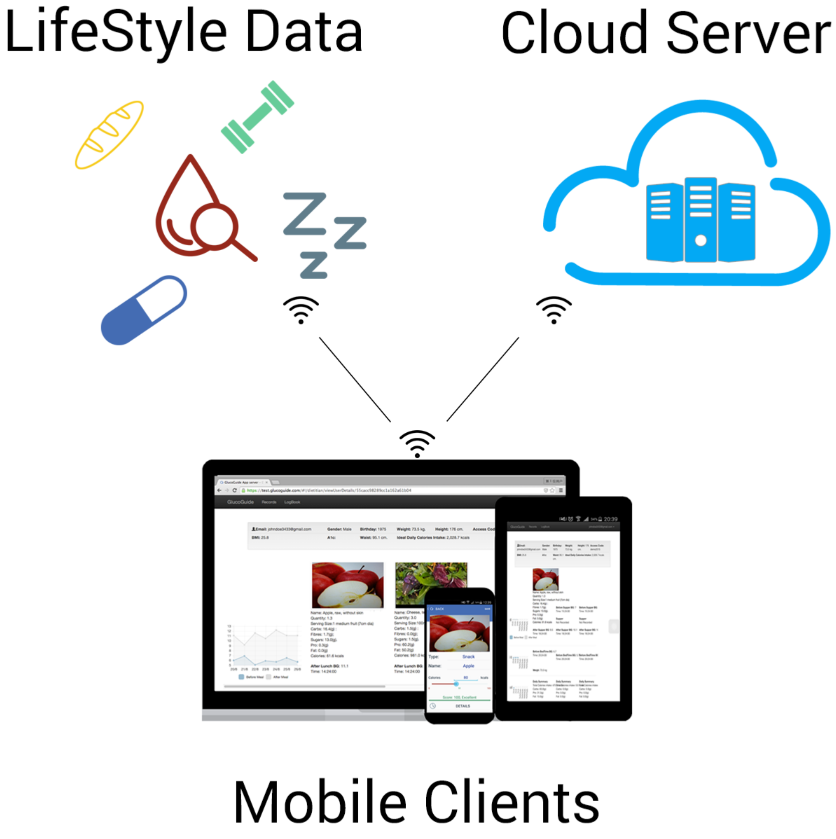


Figure 3.1: GlucoGuide architecture

The GlucoGuide Engine, consisting of various computing servers, can also be regarded as a computer cloud. It consists of a centralized database, machine learning engines, and a web portal, which allows patients and our health teams to log-in to review the visualized health data and online logbook. The core function of GlucoGuide Engine is to analyze the uploaded data and generate recommendations based on patients' data using predictive models.

3.3.1 Food Classification

The first type of recommendation is proactive diet guideline. Among all the components of self-management, the most important and complicated part is the diet management, requiring T2D patients restrict and accurately count the amount of carbohydrates, protein, fat, and so on [25]. Since there are diverse foods available at various places (supermarket, restaurant, home-made, etc.), for most T2D patients (especially for prediabetic patients) often feel frustrated and troublesome in composing effective and diabetic-friendly meal plans to achieve their daily carbohydrates goals. They do need assistance to help them determine what kinds of food items are suitable for their diabetic conditions or not. Ideally, patients should be able to query concerning the recommendation of any food items in nearly real-time.

We model this problem as a binary or multinomial classification problem, predictive models for classification thus can be employed to categorize any food items into its corresponding recommendation level (such as this food item should be chosen more often or chosen less often).

To achieve this goal, we first construct a comprehensive food feature vector via extracting nutrient and textual features from training food datasets. Then, three food recommendation labels: "*Choose More Often*", "*In Moderate*", and "*Choose Less Often*" are designed as class labels. Next, we explore and evaluate several state-of-art predictive models with different biases and structures for the food classification task. According to the results of our empirical study, we find that tree-based classification models outperform others in this practical problem context. The best prediction accuracy of tree-based models could reach to around 95%, which is a very promising result indicating the proposed idea of using classification model for diabetic food classification is empirical effective. See Chapter 4 for details.

3.3.2 Lifestyle Recommendation

The second type of recommendation is related to lifestyle activities (such as diet, exercise, etc.) and derived from the outcomes of blood glucose prediction.

If the updated lifestyle data indicate that a patient is in emergent or dangerous situations, such as abnormally high or low blood glucose levels for longer period of times, GlucoGuide provides immediate assistance to the patient, as well as alerts to the healthcare providers.

If patients' lifestyle data do not present emergent or dangerous situations as above, those data will be accumulated. A few times in a week, a proposed predictive framework called temporal-weighted regression (TWR) will be deployed on our server to discover correlations between the recent lifestyle data and the blood glucose levels for each patient. Such correlations will be framed in natural language templates (aligned with CDA guidelines) and sent to patients' mobile devices as recommendations. As patients' data are different, the recommendations are also personalized.

To evaluate the performance and the clinical effectiveness of this framework, we conducted a three-month clinical trial on human subjects. Due to the high cost and complexity of trials on human, a small but representative subject group was involved. We empirically evaluated the TWR framework and showed that the prediction performance of TWR is similar to the state-of-the-art [16], using the collected lifestyle data and discrete blood glucose samples. Furthermore, two standard laboratory blood tests for diabetes were conducted on patients before and after the trial. The results were quite remarkable with over 90% confidence levels in the significance test. In sum, GlucoGuide could be amounted to turning an early diabetic patient to be pre-diabetic, and pre-diabetic to non-diabetic, after a three-month trial. See Chapter 5 for details.

Chapter 4

Food Classification for Diabetes

Effective diet management requires T2D patients restrict and accurately count the amount of carbohydrates, protein, fat, and so on. T2D patients (especially for prediabetic) often feel frustrated and troublesome in composing efficient and diabetic-friendly meal plans to achieve their daily carbohydrates goals.

Most diet management tools for T2D patients only provide basic meal recording and nutrient visualization functionalities, such as food database searching, calories counting, etc. The primary barrier of these tools is the lack of diet guideline/assistance for T2D patients, which is typically provided by dietitians with a costly (either patients' own time costs and/or government/insurance company medical costs) and time-consuming consultation process. In fact, T2D patients do need such assistance in a more convenient and timely manner to help them decide what kinds of food items are suitable for their diabetic conditions or not, especially for those home-made food items.

As far as we are concerned, predictive models can be utilized to overcome this challenge and proactively guide patients in their food selection. More specifically, they can be employed to determine whether a food item should be chosen more often or less more.

We define this problem as a multinomial classification task. As for this problem, we first construct a comprehensive food feature vector and extract nutrient and textual information from two different training food databases. Then, three food recommendation labels: "*Choose More Often*", "*In Moderate*", and "*Choose Less Often*" are designed as class labels.

Next, we explore and evaluate several state-of-art predictive models with different biases and structures for the food classification task. According to the results of our empirical study, we find that tree-based models outperform others in this practical problem context. The best prediction accuracy of RandomForest reaches around 95%, which is a very positive result indicating the proposed idea of using classification models is promising. Furthermore, we have also explored the feature importance and instance proximity matrix of generated RandomForest.

More details about this empirical study can be found in Section 4.3.

We have deployed the fine-tuned classification model in the GlucoGuide mobile clients, along with a comprehensive food database and food UPC scanning technologies. As such, T2D patients could ubiquitously and conveniently access any food items in the database with the help of model-generating food recommendation labels.

The rest of this chapter is organized as follows: Section 4.1 describes the design principle of the proposed food classification tool, Section 4.2 presents the construction of food feature vector, Section 4.3 describes the classification models selection, evaluation methodology, and its results. The last two sections discuss the relationships to previous works and conclusions.

4.1 Design Principle

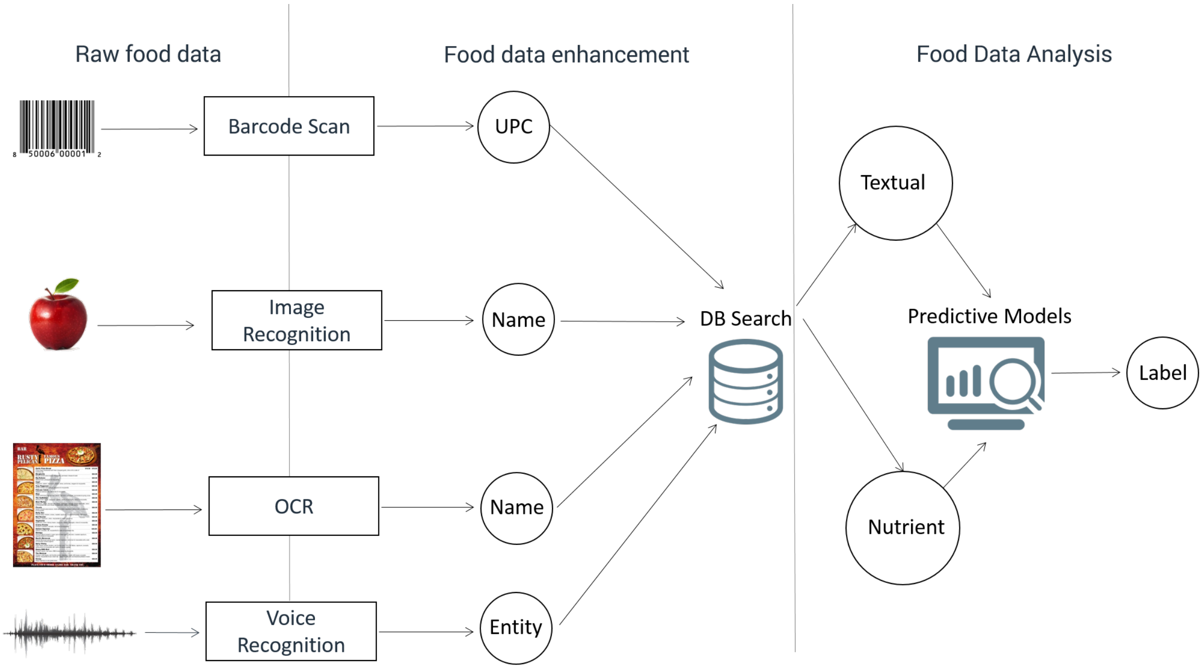


Figure 4.1: Patients nowadays can easily use mobile devices to get the detailed information about food items (nutrient and textual information). Then, our classification model can generate recommendation labels on each food item.

The design principle of our food classification tool is shown in Figure 4.1. With the assistance of many advanced HCI technologies, patients nowadays can use their mobile devices to first obtain the meta-data of food items (UPC code, entities, keyphrases, categories, etc). Then, we can use the meta-data to query our food database to get the specific nutrient and textual information about the food items. By utilizing well-trained classification models, our tool can output recommendation labels for users.

Numerous food databases are available in the health and nutrition industry. They typically have different characteristics and scales in terms of dimension and volume. For example, Canadian Diabetes Association (CDA) maintains a food database with only three nutrition features while a large scientific food database provided by Canadian Nutrition File (CNF) has more than 150 nutrition features. However, most of these food databases do not provide food recommendation to T2D patients.

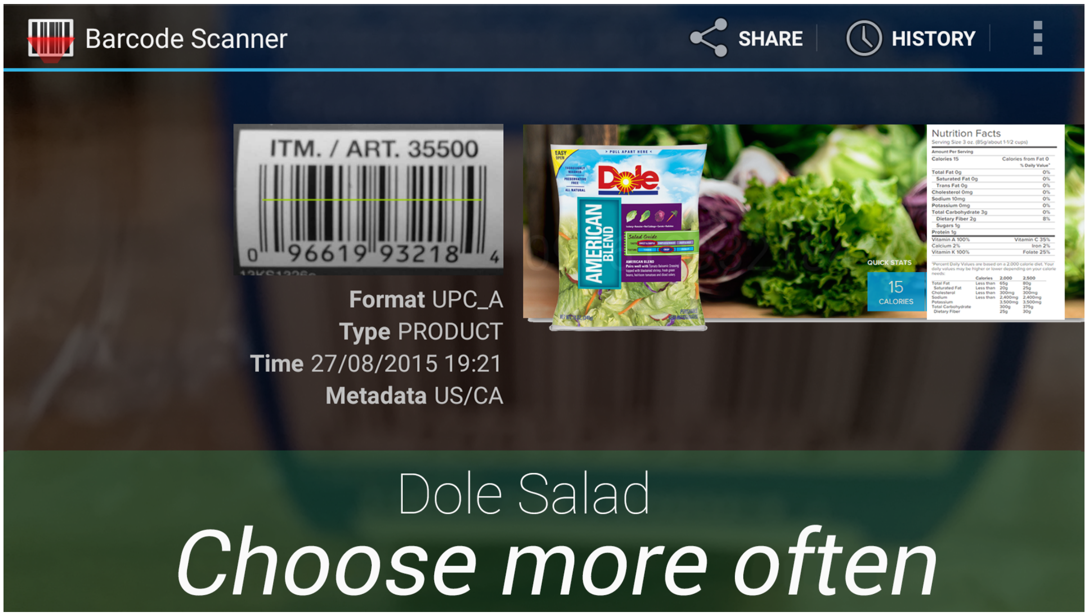
The food database used in our tool contains 65,000+ popular North-American food items (a commercialized food database), and it is a very comprehensive database with each item has its food name, category, sub-category, and 18 nutrient attributes. Also, this is the target dataset for the food classification task. Patients can search this food database to record detailed information about their diet. To facilitate the database search, we designed several interactions for users to input food keyword. They can type the keywords (such as “bagel“ or “starbucks coffee“) using a soft keyboard or using the voice dictation to speak out the keywords. More conveniently, we employed barcode scan technologies to help patients scan the food UPC code (Figure 4.2) and retrieve food nutrition and textual information.

Our training datasets are two diabetic food databases, each food item in the databases has been labeled by dietitians and nutrition experts. In particular, one training dataset is provided by CDA [2] containing around 500 most common food items in Canada. CDA classifies each food item to be “*Choose More Often*“ or “*Choose Less Often*“ (e.g., binary labels) as a food guidance for diabetic patients. Another one is provided by Food Picker $^{\text{TM}}$ containing more than 13,000 food items. Each food item has three class labels: “*Choose More Often*“ , “*In Moderation*“, or “*Choose Less Often*“ (e.g., muti-nominal labels). Note that these two databases are collected from two different data sources, with different features, food items, and labeling strategies. Thus, the classification model to be learned could absorb more comprehensive domain knowledge via such data aggregation.

Next, we construct the food feature vector for classification. Ideally, a large feature vector is desirable to characterize food items comprehensively. Based on the properties of food, we defined two types of features: nutrition features and textual features.

In particular, nutrient features characterize a food item via its nutrition attributes, such as carbohydrate, fat, protein, sodium, etc., which are costly but accurately calculated by numerous nutrition laboratories worldwide. Most food databases provide nutrient information by default. The second feature is related to a food item’s textual information, such as category, subcategory, name, etc. This kind of feature can also be very useful for classification especially when the nutrition information is not available. How the textual features can be extracted and integrated will be described in Section 4.2.

Once the training datasets are prepared, we then need to select the classification model(s) for the classification task via empirical evaluations. The best fit model will be deployed on GlucoGuide mobile clients to classify food items.



Format UPC_A
Type PRODUCT
Time 27/08/2015 19:21
Metadata US/CA

Dole Salad
Choose more often

Figure 4.2: Scan the barcode on a salad product to get its UPC code, then use the UPC code to retrieve nutrition and textual information from UPC database. Our classification model tries to classify this food item based on the nutrition and textual information.

With this tool, patients can conveniently track and review their daily nutrition intakes. More intelligently, the embedded model can guide patients in their food selection. Here is an example illustrating how T2D patients could use this tool. Suppose a patient is purchasing groceries in a supermarket and having difficulties in choosing a food product. He/she can then open our tool to scan the barcode of the product (such as Dole Salad) and get its nutrient and textual information. Then, our model can output the color-coded class label to guide the patient about this food, as shown in Figure 4.2.

4.2 Textual Feature Extraction

We believe that a food item can not only be classified by its the nutrient features but also by its textual features as supportive information. As such, we define the full feature vector of a food data instance as:

$$F = [\mathcal{N}, \mathcal{T}]$$

where $\mathcal{N}$ represents the nutrition feature vector and $\mathcal{T}$ represents the textual feature vector. Since the values of nutrition feature $\mathcal{N}$ are already known in both training food databases, how to extract textual features and generate $\mathcal{T}$ is the main component of our feature extraction task.

In particular, nutrition features are obtained straightforward from the nutrition label. Once we have obtained the nutrition features of food items, we can utilize some state-of-art classification models without taking any domain knowledge into consideration. On the contrary, constructing the textual features do require domain knowledge and specifically-designed feature extraction strategies.

We choose to extract textual features mainly from the CDA food database since it is an official food database [2]. In other words, it provides the most authoritative textual information.

Given the food vocabulary vector V of CDA's training food database, each dimension corresponds to phrase w_t from V. Typically, the dimension of the vocabulary is large. To reduce the dimension size, we replace it with a subset of keyphrases vector K. Note that we only use 1-word or 2-word keyphrases such as "whole wheat", "low fat", and "sweeten", etc., appear in the food name.

The next step becomes how to extract the important keyphrases existing in the food vocabulary vector V. We proposed two approaches to generate the keyphrases vector K to achieve this goal.

4.2.1 Domain Knowledge based Extraction

Human labeling is an effective approach to extract keyphrases, although it could be very time-consuming and expensive. Thus, with the assistance of our medical team, we identified a human-labeled keyphrase list $k_h, h \in \{1, \ldots, H\}$ via thoroughly and delicately searching in CDA literature and its official diet guideline [2].

However, the human extracted keyphrase list could not be comprehensive enough. To enrich the human-labelled keyphrase list K_h, machine learning based feature extraction will also be used.

4.2.2 Machine Learning based Extraction

We have employed the NaïveBayes learning to extract the keyphrases, which is widely used in document classification. The influence of each phrase in a food item was measured with respect to its class to discover the most important/informative keyphrases.

The training food items in CDA database provide binary class labels. To simplify the format, we define the class “*Choose More Often*“ as c_0 and “*Choose Less Often*“ as c_1. We treat each food item i as a document d_i, which is a binary vector over the vocabulary V. Then, we characterize the probability of food item with its class using multi-variate Bernoulli event model [38, 39] (following the Naïve Bayes assumption):

$$P(d_i|c_j) = \prod_{t=1}^{|V|}(B_{it}P(w_t|c_j) + (1 - B_{it})(1 - P(w_t|c_j))). \tag{4.1}$$

where $B_{it} \in \{0, 1\}$ indicates whether phrase w_t appears at the dimension t of food item d_i. Then, we build a NaïveBayes classifier attempting to find the class c_j that maximize the probability $P(d_i|c_j)$.

Given the CDA training data $D = \{d_1, d_2, \ldots, d_{|D|}\}$, we calculate the binary-class conditional phrase probability as:

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{it}P(c_j|d_i)}{2 + \sum_{i=1}^{|D|} P(c_j|d_i)}. \tag{4.2}$$

Note that a Laplacean prior in Equation 4.2 is calculated in case some phrase counts are zero. Having such phrase probability, we define a keyphrase selection metric Q_t as:

$$Q_t = \frac{P(w_t|c_0)}{P(w_t|c_1)}. \tag{4.3}$$

By intuition, Q_t indicates how likely the phrase w_t associated with class c_0 compared to class c_1. In another phrase, any phrase with a large Q_t value means it is a class discriminative keyphrase with respect to classification. We thus choose phrase w_t as a candidate keyphrase k_m (note that each keyphrase here contains only one phrase) if its Q_t is beyond a predefined threshold (5th quintile was chosen in this experiment). By iteratively performing the above calculations, we can build a machine-generated keyphrase list $k_m, m \in \{0, \ldots, M\}$.

4.2.3 Merge Textual Features

The last step is to append the extracted keyphrase vector K_m list to the existing human labeled keyphrase list K_h. We first compare each keyphrase k_m with all human labeled keyphrase K_h. If a keyphrase k_m is “similar“ enough to an existing keyphrase k_h, we just ignore the keyphrase and move to the next one on the list. Here we employ a metric $Q(k_m, k_h)$ defining such similarity between k_m and keyphrase k_h:

$$Q(k_m, k_h) = 1 - \frac{L(k_m, k_h)}{l_{max}(k_m, k_h)}, 0 \leq Q(k_m, k_h) \leq 1. \tag{4.4}$$

where $L(w, k)$ is the Levensthein distance [40, 41] between k_m and k_h. The $l_{max}(k_m, k_h)$ is the maximum string length between k_m and k_h. As such, for each phrase k_m, we calculate the maximal similarity $Q(k_m, k_h)_{max}, h \in \{1, \ldots, H\}$. If $Q(k_m, k_h)_{max}$ is below a predefined threshold, we will consider k_m as a newly discovered keyphrase and append it to the human-labelled keyphrase list K_h.

4.3 Generating the Training Dataset

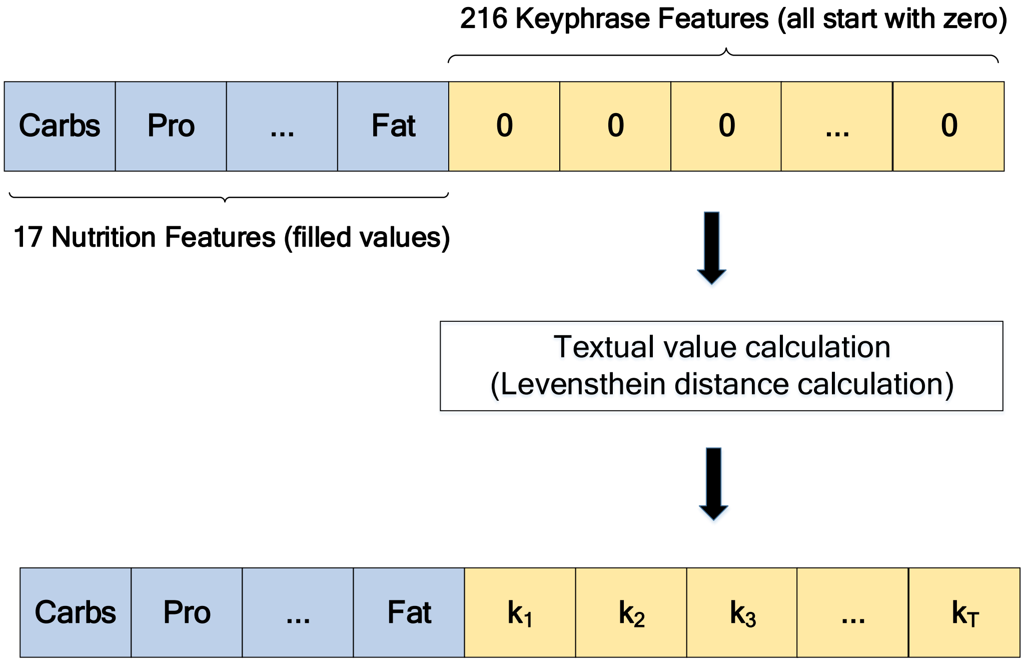

216 Keyphrase Features (all start with zero)
Carbs
Pro
...
Fat
0
0
0
...
0
17 Nutrition Features (filled values)
Textual value calculation
(Levensthein distance calculation)
Carbs
Pro
...
Fat
k1
k2
k3
...
kT

Figure 4.3: Aggregating the nutrition features and textual features, and calculating the values of textual features using Levenshtein distance.

After merging the two keyphrase lists K_h and K_m, we can generate the textual feature vector $\mathcal{T}$. For each food item in the training databases, we combine the textual feature $\mathcal{T}$ to the

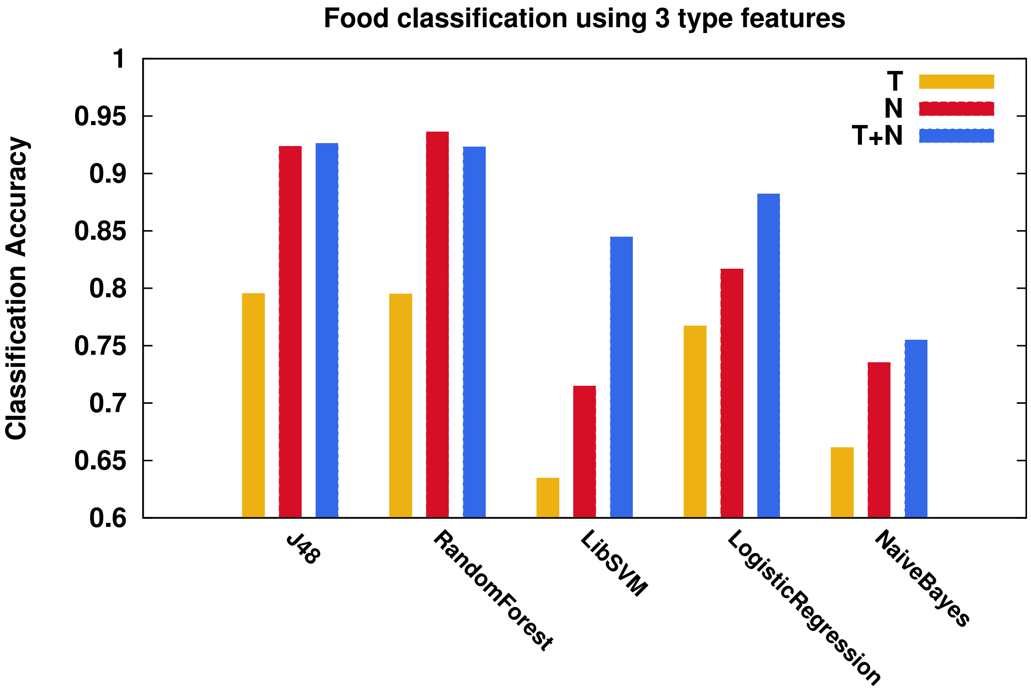


Figure 4.4: Classification accuracy for each model using 3 different feature vectors

nutrition feature vector $\mathcal{N}$ and obtain the complete structure of feature vector F, as shown in Figure 4.3.

The complete feature vector F has 17 nutrient features and 216 textual features; each nutrition feature $\mathcal{N}_i$ represents a nutrient element and each textual feature $\mathcal{T}_j$, $(0 \leq \mathcal{T}_j \leq 1)$ represents the keyphrase similarity. The value of $\mathcal{T}_j$ is calculated by measuring the maximal similarity between each keyphrase feature $\mathcal{T}_j$ and all the subsequences whose length less than three words of the food name. Such calculation process is similar to Equation 4.4.

After the feature extraction and data preprocessing, the training dataset (including 233 features and 13,260 data instances) is ready for the classification task.

In this section, we will discuss the classification model selection and evaluation.

4.3.1 Evaluation Methodology

4.4 Model Selection and Evaluation

Different classification models have different cost functions, structures, and biases. As far as we are concerned, choosing the best fit model for the food classification task is an iterative and experimental process. Given the characteristics of this food classification problem, we have chosen five popular classification models as candidates: *J48* [42], *RandomForest* [43], *LibSVM* [44], *NaïveBayes* [45], and *LogisticRegression* [45].

Next we will describe the experimental settings for these classification models. Finding the optimal parameters for each model is a necessary but time-consuming process. For each model, we conduct ten fold cross-validation first with different parameter settings to ensure it could reach the potential on this dataset. For example, all tree-based models were pruned with confidence level set to 0.25 [42], the number of trees in the RandomForest was set to 10, the kernel of LibSVM was set to the RBF kernel, the standard classification *accuracy* was selected as the main performance measurement. etc.

In addition, we evaluate each model using three different feature sets (i.e., textual features $\mathcal{T}$, nutrition features $\mathcal{N}$, and comprehensive features ($\mathcal{T} + \mathcal{N}$)). As such, we can empirically evaluate the different combinations of feature vectors and classification models then find the best one.

After the ten folder cross-validation with optimal parameter tuning, we obtain the model accuracies, as shown in Figure 4.4. We can see that all models have the worst classification accuracy using only the textual feature vector (yellow bars) and have much better accuracy using the nutrition feature vector (red bars). Furthermore, by integrating the textual and nutrition feature vectors, the accuracy is significantly improved for almost all models except for RandomForest (blue bars). The classification precision and recall have the similar trend as the overall accuracy thus we do not depict them in here.

Note that J48 and RandomForest outstand from other candidates on this dataset. They are able to obtain around 95% prediction accuracy using the combination of nutrition and textual features (Figure 4.4). One possible explanation could be that the tree models are the most suitable model for mimicking the dietitians' reasoning/classification process. Also according to Kotsiantis [46], tree models are sequential models, which logically combine a sequence of simple tests; each test compares a numeric attribute against a threshold value or a nominal attribute against a set of possible values. In that way, they can closely resemble human reasoning and are easy to understand. As a result, the tree models have been discovered to be the best model for the food classification task in this problem domain.

Also in this experiment, we find that textual features based food classification is also promising, as it can obtain about 80% classification accuracy using tree models. Thus, we believe that textual features based classification could be useful especially when the nutrition features are not accessible or available. For example, if the nutrition information of a food item is unknown, patients can still use modern HCI technologies such as barcode scan or OCR to obtain the textual information (name, categories, brand names, etc.). Tree models can still conduct classification tasks on the text-only test instances with reasonable good prediction accuracy.

In sum, in this empirical study, we find that RandomForest model is the best-fit model among all the candidates. Thus, we can conduct further empirical studies using the RandomForest model.

4.5 Food Classification Using RandomForest

RandomForest [43] is a very powerful and comprehensive classification model. It is a combination of randomly generated classification trees. For each test instance, each tree in the forest generates a class label and the final class label is the one gets the most "vote". RandomForest has many advantages, including the generation of an internal unbiased estimate of generalization error (out-of-bag error), feature importance estimation, case proximities, and handling large scale of data, etc.

4.5.1 Model Tuning

In RandomForest, each tree is generated using a different bootstrap sample from the original data sample. About one-third of the cases are left out of the bootstrap sample and not used in the tree construction as the test set. In this way, a test set classification is obtained for each case in about one-third of the trees. The proportion of times the estimated class is not equal to the true class averaged over all cases is called the out-of-bag (OOB) error estimate. As such, RandomForest typically dose not require cross-validation or separate test datasets to get unbiased performance estimation [43].

RandomForest has two importance parameters to be tuned. The first one is the number of features randomly sampled as candidates at each split, and the second parameter is the number of the trees in the forest.

We used grid searching to tune the first parameter, i.e., the number of feature candidates, started with 15. We set the inflation/deflation rate to 1.5. For each tuning iteration, the OOB error estimation is compared with the previous iteration until converged (improvement less than 0.05). In this experimental context, we find that the optimized feature number is 73.

As for tuning the number of trees in the forest, we investigated the OOB with 50, 100, 250, 500, 1000 trees, as shown in Figure 4.5. The feature candidate number was set to 73, as discussed above. As we can observe from Figure 4.5, the OOB error rate converges at 100 trees. Thus, we can choose the optimized tree number to be 100 for the future experiments.

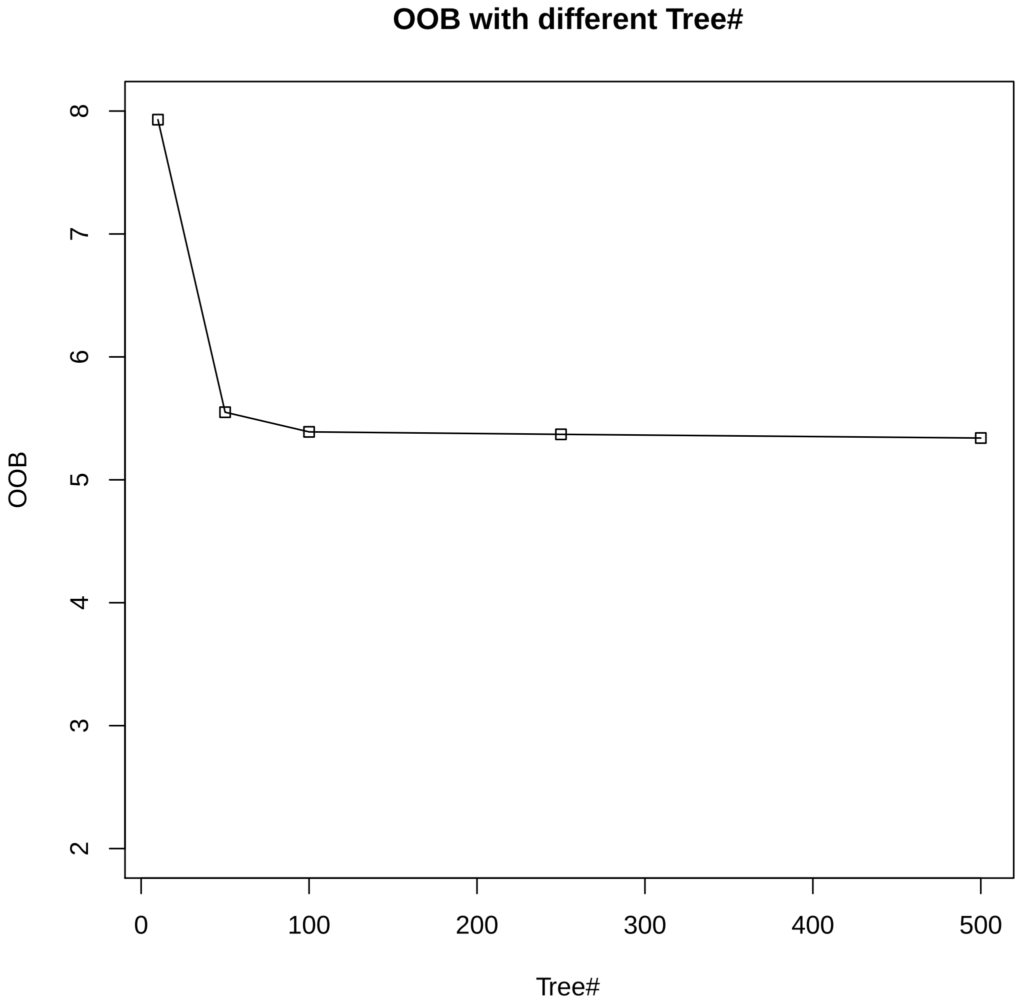


Figure 4.5: OOB error rate with different tree numbers in the forest

With the fine-tuned RandomForest, the classification error rate is 5.37%, meaning a very promising 94.63% classification accuracy. To further estimate the classification error rate for each class, we output its confusion matrix, as shown in Table 4.1. We can see from the matrix that although the class distribution of original data sample is not well balanced, the estimation error rates for each class are still similar. As such, we think no imbalanced classification technologies such as cost matrix or class weighting are required for this food classification task.

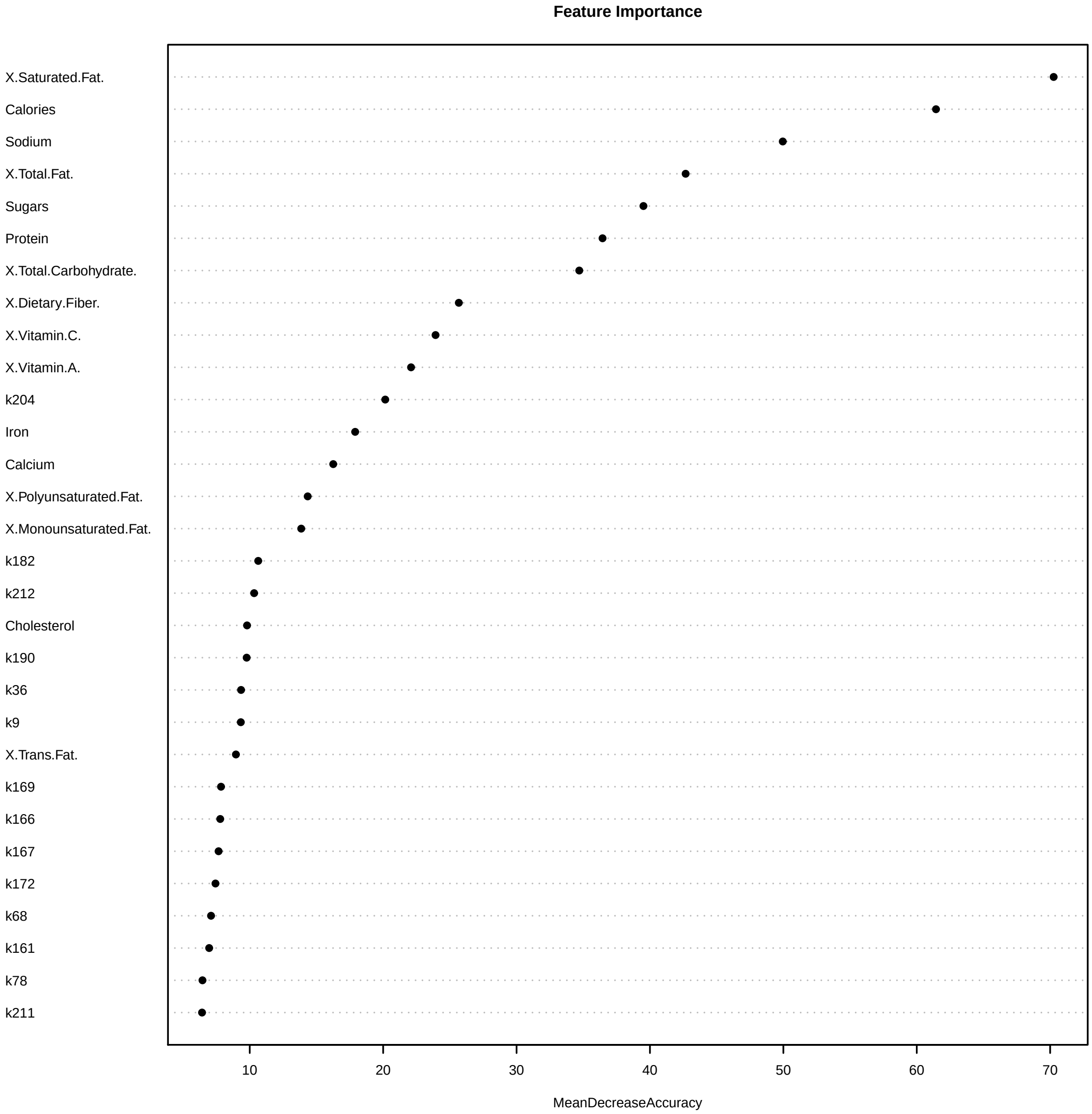


Figure 4.6: Feature importance with respect to mean accuracy decrease.

	In Moderation	Less Often	More Often	class error
In Moderation	6317	282	50	0.04993232
Less Often	336	5277	8	0.06119907
More Often	33	5	952	0.03838384

Table 4.1: Classification confusion matrix

4.5.2 Feature Importance

The dimension of our food feature vector is 233. To increase the interpretation, we want to discover which features have the large influence on prediction results. In RandomForest, the feature importance of feature m is calculated by first randomly permute its values in the OOB cases and put these cases down each tree in the forest. Then we subtracted the number of votes for the correct class in the variable-m-permuted OOB test cases from the number of votes for the correct class in the untouched OOB cases. The average of this accuracy decrease over all trees in the forest is the raw importance score for the variable m [43]. We depict the top important features in Figure 4.6, we can see that among all the features, the most importance nutrient features are: Saturated Fat, Calories, Sodium, Total Fat, Sugars, Protein, Carbohydrate, and Fiber. The most discriminative 1-word keyphrases are “wheat“, “light“, “grains“, “fruit“, “milk“, “sugar“, “cheese“, “diet“, “dressing“, “yogurt“, “unsweeten“, “free“, etc.

Note that the important features we discovered in this empirical study are consistent with the nutrient and diet guideline provided by CDA [25] and Health Canada. This finding also validates the effectiveness of our classification model.

4.5.3 Prototype

RandomForest can also be used to calculate a N by N (N is the total number of data instances) proximities matrix [43]. After a classification tree is built, we can put all of the data instances down the tree. If cases k and j are in the same leaf node, their proximity in the matrix will be increased by one. The final proximity matrix is normalized by the total number of trees in the forest.

After the proximity matrix is calculated, we can further obtain the “representative“ instances on each class label. Such instances are called Prototype [43], which is a similar concept to centroid in clustering but generated in a supervised manner.

We calculate food item prototypes (nutrient features only since they are much more important) with respect to each class, as shown in Table 4.2. We can see that some features such

	Choose More Often	In Moderation	Choose Less Often
Calories Total (g)	5	110	270
Total.Fat. (g)	0	1	14
Saturated.Fat (g)	0	0	6
Cholesterol (mg)	0	0	30
Sodium (mg)	20	380	460
Carbohydrate (g)	1	20	27
Dietary.Fiber (g)	0	3	2
Sugars (g)	0	2	5
Protein (g)	0	0	8
Vitamin.A (%)	0	0	6
Calcium (%)	0	4	10
Iron (%)	0	8	6

Table 4.2: Food classification prototypes

as Calories, Sugars, Carbohydrate, Sodium, and Total Fat increase monotonically with class labels.

Other features such as Fiber and Iron do not follow such pattern. This inconsistency could be caused by low measurement granularity. Any value less than 1g or 1% will be given zero value in the data instance. Thus, the fiber value for "*Choose More Often*" prototype is in fact unknown due to the granularity. However, if we output the fiber calories breakdown ratio in total carbohydrates, we can see that the ratios are 0%, 15%, and 0.07% respectively. Still, it is reasonable to assume that in healthy foods, the percentage of fibers in total carbohydrates should be much larger than the unhealthy foods. We think that the fiber percentage possibly also increase monotonically with class labels.

In sum, features like fiber percentages, as well as other calories breakdown ratios, could be added in the food feature vector as new features to improve the classification accuracy. We could collaborate with nutrition experts to discover more such features. More food feature engineering will be one of our future work.

4.6 Relationship to Previous Work

Food labeling/classification for T2D diabetes is an interdisciplinary area and has been researching for several decades. In this area, the glycemic index (GI) is the most impact food labeling system created on 1980 [47].

More specifically, GI is a measurement of the power of foods to raise blood sugar glucose levels after being eaten. Food items with a high GI value contain rapidly digested carbohydrate, which causes a rapid rise and fall in the level of blood glucose. In contrast, food items with a low GI value contain slowly digested carbohydrate, which causes a gradual, relatively low rise in the level of blood glucose. The GI values of foods must be measured using valid and expensive scientific methods [47].

Sydney University GI Research Service (SUGiRS) provides a reliable commercial GI testing laboratory for the local and international food industry [48]. They created GI values of more than 2500 foods. However, as far as we are concerned, GI system is still too complicated, expensive, and not user-friendly for general T2D diabetes.

Guiding Star [49] system is another nutrition guidance system aiming to score the food items in Canada's most popular supermarkets manually. Its food labeling is a collaborative process between grocery retailers and a panel of nutrition experts. However, as far as we are concerned, the classification rules of Guiding Star system are all fixed thus not data-driven, which makes its scalability is very low. Moreover, its guideline is designed for general people, which is different from the diet guidelines for people with T2D.

Compared to the previous works, our work proposed a more flexible and reliable framework for this problem, which makes the following contributions. Firstly, the concept itself is novel, to the best of our knowledge, this work is the first attempt integrating machine learning technologies into T2D patients' diet management. The second contribution is that we have designed an effective approach to extract a comprehensive food feature vector, including both nutrient and textual features, for the classification task. The third contribution is that we find that the tree-based classification models (J48, RandomForest, etc.) can archive the best performance in this problem context. We also evaluate the performance of tree-based models using the different feature and instance sets. Furthermore, the optimal features are discovered after the evaluation. The last contribution is that the deployment of the classification model on GlucoGuide mobile clients, allowing patients to access food database and the embedded classification model conveniently and ubiquitously.

4.7 Food Classification Future Works

Even with the help of barcode scanning and food database searching, food recording could still be difficult. Ideally, when a meal is taken pictures by the camera of patients' mobile devices, labels and nutrients of this meal are automatically analyzed by images pattern recognition algorithms. Such automatically food recording process would greatly enhance the usability and user experience. However, image-based food information retrieval is extremely challenging in generic objects recognition due to its large varieties in spatial, shape, color, texture features, etc. Due to these complex characteristics, traditional multinomial classification solutions would cause large intra-class errors in practical settings.

Deep learning frameworks could be applied to classify food items based on food photos. In fact, the basic ideas of deep learning appeared about two decades ago. However, due the limitation of computational power, modeling a deep Artificial Neural Networks (ANN) would be extremely time-consuming (normally with the unit of years). As such, researchers at that time only applies “shallow“ ANN but still achieve state-of-art results. In 2011, Andrew Ng founded the Google Brain project at Google, which developed deep ANN using Google’s distributed compute infrastructure. Among its notable results was a neural network trained using deep learning algorithms on 16,000 CPU cores, that learned to recognize higher-level concepts, such as cats, after watching only YouTube videos, and without ever having been told what a ”cat” is. The project’s technology is currently also used in the Android Operating System’s speech recognition system.

Although deep learning can automatically learn the hierarchical feature representations and achieve very promising precision, it normally needs a huge amount of training data. In another word, if data are limited, the performance of deep learning might not be able to outperform some simple learning algorithms. Also, in deep learning, the time complexity is also big, to assure the feasibility of the algorithm, high-quality distributed computing and hardware supports are required.

The computer vision team of GlucoGuide has been researching and developing mobile camera-based food recognition for more than two years now. We have applied Convolutional Neural Networks (CNN), a state-of-art deep learning approach, to recognize the food item through parameter optimization on GPU. We constructed a dataset composed of 49,360 food images of 40 types most popular Canadian and American food and trained a Convolutional Neural Network classifier for food recognition. Our test results show promising accuracy rates (top 1 accuracy rate: 78%; top-five accuracy rate: 93%). With the aid of food recognition technique and high-end GPU computers, users are able to conveniently log their diet data just by snapping a photo and getting an estimate of the nutrition information in less than one second.

4.8 Conclusion

In this chapter, an intelligent food classification tool is proposed using classification models and mobile computing, aiming to facilitate T2D patients’ diet management and provide food recommendations. We have collected two training food database with true labels and extracted a comprehensive feature vector from them. Extensive empirical studies have been conducted to explore among different types of classification models. The studies indicate that the RandomForest is the best-fit model with around 95% prediction accuracy. The well-trained model is also deployed on GlucoGuide mobile clients to help T2D patients conveniently manage their diet.

This tool has proved empirically effective to provide proactive food guidelines allowing T2D patients to make decisions on their food selection beforehand. In addition to food recommendation, we also want to provide recommendations regarding patients' historical lifestyle activities. To achieve this goal, we have designed and implemented a lifestyle recommendation framework to generate data-driven recommendations helping T2D patients to stabilize their blood glucose. We also conduct a clinical trial to collect training data and evaluate its clinical effectiveness. The details of this framework and its evaluation results will be presented in the next chapter.

Chapter 5

Lifestyle Recommendation Framework

In this chapter, we will discuss how the comprehensive lifestyle recommendations, are generated and distributed.

The proposed lifestyle recommendation framework is generally illustrated in Figure 5.1. It contains three main components: 1) *Lifestyle Data Aggregation and Preprocessing*, 2) *Temporal-Weighted Regression (TWR)*, and 3) *Model Postprocessing*.

To evaluate this framework, we conduct a 3-month clinical trial to collect training data and generate recommendations. Patients during the clinical trial upload various types of data simultaneously or at any times of a day; some are very noisy and abundant. As such, a data aggregation and preprocessing module is designed to convert the raw lifestyle data into proper training datasets for predictive tasks. Then, the proposed TWR (Temporal-Weighted Regression) should be able to identify the temporal patterns of blood glucose levels and convert them into computer-generated recommendations.

In next sections, each component of this framework will be described in detail.

5.0.1 Data Aggregation

The GlucoGuide mobile client can collect more than fifty health features (such as medicine records, meter readings, nutrients, physical activities, water intake, sleep pattern, weather, and so on) for each patient daily. It is a very comprehensive feature vector which almost cover every aspect of T2D diabetes management. However, it might not be feasible to include all these features in the model, due to the data sparsity, missing values, noise in the data, and the feature irrelevance. Thus, finalized by our medical team, the most relevant sixteen features were delicately selected to characterize the main characteristics of T2D patients' daily activities, as listed in Table 5.1.

All heath features except for nutrient intake (discussed in the previous section) were recorded

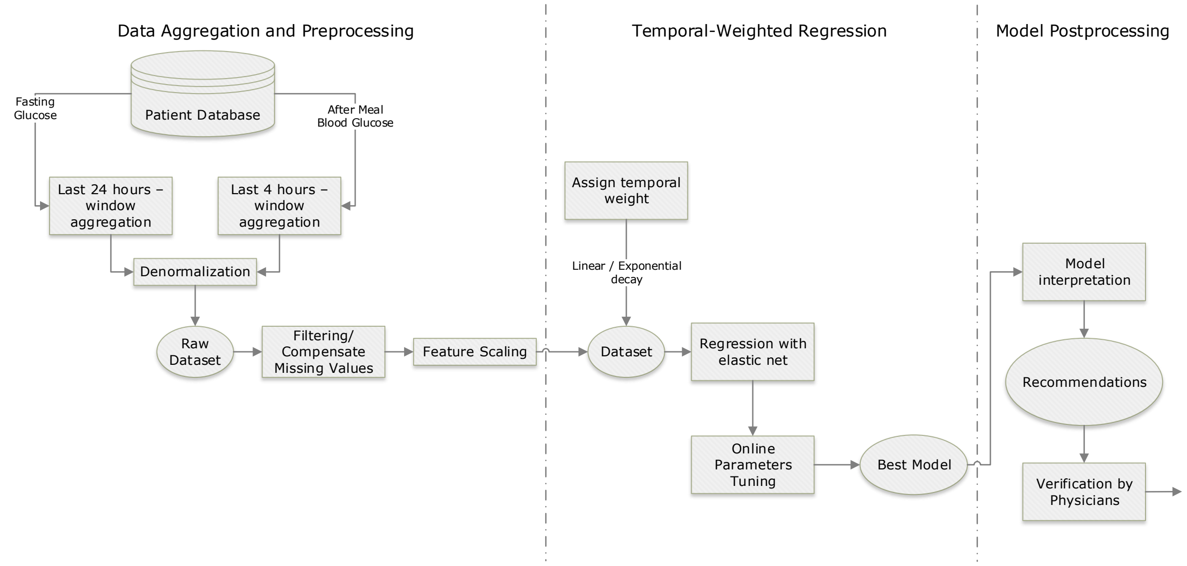

Data Aggregation and Preprocessing
Temporal-Weighted Regression
Model Postprocessing
Patient Database
Fasting Glucose
After Meal Blood Glucose
Last 24 hours – window aggregation
Last 4 hours – window aggregation
Denormalization
Raw Dataset
Filtering/ Compensate Missing Values
Feature Scaling
Assign temporal weight
Linear / Exponential decay
Dataset
Regression with elastic net
Online Parameters Tuning
Best Model
Model interpretation
Recommendations
Verification by Physicians

Figure 5.1: T2D lifestyle recommendations framework

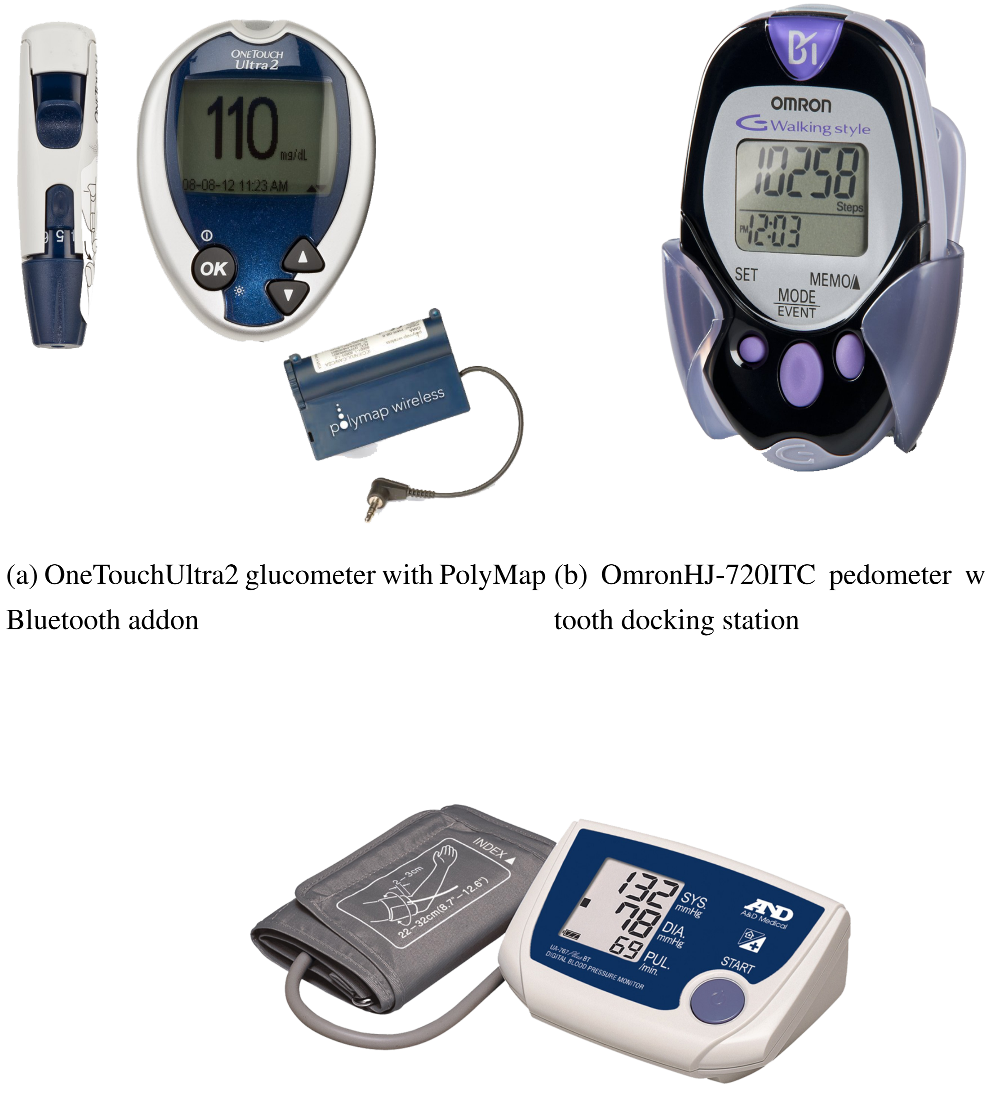
via different medical devices as described in Figure 5.2. For example, the glucometer (OneTouchUltra2, see Figure 5.2a) can record the blood glucose level and automatically send the readings to GlucoGuide clients via Bluetooth. The full Bluetooth-enabled medical device list is shown in Figure 5.2.

Among all features listed in Table 5.1, the blood glucose level is treated as the dependent variable or target variable and the rest fifteen features are used as independent variables or features. Some previous works have shown correlations between blood glucose level and the independent variables we used [50, 51]. However, the exact relation will be found in our predictive models, to be discussed next.

5.0.2 Data Pre-processing

Not like the continuous CGM data, diabetes specialists (doctors) usually advise T2D patients to check their blood glucose level (by using glucometers) at the following two critical times of the day. The first one is the fasting blood glucose, usually taken early in the morning before breakfast. This may reflect the overall status of the diabetic condition in the previous day. The second is the blood glucose level taken two hours after a main meal. Such blood glucose levels reflect more directly the effect of the meal two hours ago.

Our health team advise our patients in the clinical trial to check their fasting blood glucose levels daily, and their after-meal blood glucose regularly (mainly the main meal of the a day).



(a) OneTouchUltra2 glucometer with PolyMap Bluetooth addon

(b) OmronHJ-720ITC pedometer with Bluetooth docking station

(c) AND UA0767BT blood pressure monitor

Figure 5.2: Medical devices list in clinical trial

Table 5.1: Health features used in clinical trial

Relevant Feature	Recording Method	Unit
Blood Glucose Level	Glucometer	mmol/L
Last Blood Glucose Level	Glucometer	mmol/L
Pulse	Blood Pressure Monitor	beats/minute
Systolic	Blood Pressure Monitor	mmHg
Diastolic	Blood Pressure Monitor	mmHg
Step Count	Pedometer	Steps
Aerobic Step Count	Pedometer	Steps
Carbohydrates Intaken	Food database	gram
Fat Intaken	Food database	gram
Proteins Intaken	Food database	gram
Calories Intaken	Food database	kcal
Calories Consumption	Estimated via step counts	kcal
Meal Counts	Counting	times
Carbohydrates Ratio	Food database	percentage
Fat Ratio	Food database	percentage
Proteins Ratio	Food database	percentage

The recorded blood glucose levels, along with other raw lifestyle data, must be pre-processed and converted into a specific format so that predictive algorithm can apply. In addition to the traditional de-normalization process, we design a practical data aggregation scheme.

Specifically, we have created two types of data instance for the predictive task:

- **Fasting instance:** When GlucoGuide Engine receives a fasting blood glucose recorded by a patient, the blood glucose timestamp is first extracted. Then, lifestyle data recorded within the window of the previous day are aggregated to generate an entry of the fasting dataset.

- **After-meal instance:** When GlucoGuide Engine receives an after-meal blood glucose recorded by a patient, the blood glucose timestamp is extracted. Then, lifestyle data within the 4-hours window before that timestamp are aggregated to generate an entry of After-meal dataset. The window is set to 4 hours to include more relevant data.

Data cleaning process is then performed to obtain qualified data instances. In particular, duplicate data are removed since they presented a source of error. Missing values are replaced with the rolling average value of each individual's data. All numerical features (except for the target label blood glucose) are scaled to be in the range of [0, 1].

Note that our clinical dataset is not publicly accessible since we must follow an approved clinical ethics protocol. However, other researchers can contact us and sign a confidentiality agreement (CA) to obtain a de-identified copy of the dataset.

To mine the lifestyle data after pre-processing, we have proposed a predictive algorithm called TWR (Temporal-Weighted Regression). It can learn timely health patterns and generate recommendations accordingly with minimal generalized errors. We will discuss this framework in the subsequent subsections.

5.0.3 Temporal-Weighted Regression (TWR)

Challenges and Contributions

Designing and adopting an efficient, robust, and reliable model to mine real T2D patients' data stream is challenging. Firstly, T2D lifestyle data is costly since they are from a variety of sensors, food database queries, real blood glucose samples from patients. The cost of lifestyle data collection indicates that the scale of such dataset could not be large and we need to overcome the overfitting issue (too many features vs. few data instances). Secondly, patients may forget to upload the data, or make mistakes during the data recording and uploading. These issues cause the data stream to be noisy and sparse. On the other hand, in order to keep patients motivated, we need to start sending recommendations shortly, within several days, instead of weeks or months later. Lastly, as more data are accumulated over time and patients have been making good lifestyle changes, our system must weigh the recent data more than those in the distant past. Clearly, all of these domain-specific challenges require the model to be robust and adaptive.

Since the dependent variable (blood glucose) and independent variables (shown in Table 5.1) are all numeric, regression framework thus becomes the most suitable solution for the blood glucose prediction task. Also, regression with regularization is well-known to prevent model overfitting, and generate models for better interpretation (smaller or zero feature coefficient).

Although regression with regularization has been researched for years [52], [53], [54], we still make the following contributions in the TWR framework.

The first contribution is that in TWR we proposed a straightforward yet effective temporal weighting mechanism, which is essentially a function of time to weigh data uploaded at different times. The more recent data, the more weight they carry. Thus, TWR adapts quickly to the change of new lifestyle data as patients are taking our recommendations. The second contribution is the implementation of an online parameter tuning strategy that automatically conducts the model and feature selection. Last but not least, the deployment of TWR into T2D treatment itself is novel and is shown to be effective in the clinical trial (see later). This research could also inspire other researchers to design better predictive applications in other mobile health domains.

At a high level, TWR has been deployed in the GlucoGuide Engine and triggered weekly to generate lifestyle recommendations for each patient. In every recommendation cycle, it generates a group of candidate models for each patient based on different parameter settings. The model with the smallest error on patient blood glucose prediction was chosen. From the best model, TWR greedily chooses the most relevant health feature(s) and converted it into recommendations using predefined NLP (Natural Language Processing) templates and domain knowledge. We will present technical details about TWR below.

TWR Cost Function

TWR is essentially a form of locally weighted regression, a memory-based framework that performs regressions around a data instance of interest using only the training data that are “local“ to that point [52]. It weights the training instances according to their distance to the test instance and performs regression analysis on the weighted data. Training instances close to the target instance receive higher weights while those far away receive low ones.

Many distance-based weight functions, such as Euclidean distance, Gaussian kernel, and so on, can be used. For our problem, we use the temporal-factor to measure such distance for two main reasons. Firstly, as we have mentioned, the more recent data instances should be considered more informative in the training process. Secondly, our clinical data tend to be noisier at the early stage of the clinical trial since the subjects are most elderly, and some of them have long learning curves about GlucoGuide system.

Suppose we have collected N training instances for a specific patient. Each instance has a target blood glucose level as the dependent variable y_i, data entry x_i with p dimensions, and coefficient vector θ. We assume that the relationship between the independent variables (health features) and dependent variable (blood glucose) is linear. According to Zecchin et.al [20], the performance of linear and complex nonlinear models (such as ANN) on CGM blood glucose predictions are not significantly different. However, linear model has the advantage of simple structure, fast learning rate, and high interpolation. Thus, we first explore the usage of the linear model on blood glucose prediction.

We define the cost function for the blood glucose regression model as in Equation 5.1.

$$J(\theta) = \sum_{i=1}^{N} w_t^i (y_i - \theta_0 - x_i \theta^T)^2. \tag{5.1}$$

Here we assign a dynamic weight w_t^i following a temporal decay for each training instance i. Two typical decay functions are adopted to characterize the temporal decay: *linear decay function* and *exponential decay function*. More specially, given an instance i, its weight w_t^i is defined as:

$$w_t^i = \begin{cases} 1 - \frac{d_l - d_i}{L} & \textit{Linear,} \\ \\ \frac{\beta^{(d_l - d_i)}}{\sum_{j=0}^{l} \beta^{(d_l - d_j)}} & \textit{Exponential.} \end{cases} \tag{5.2}$$

where d_l is the date when recommendation is generated (triggered) and d_i is the date of instance i. Also, L determines the linear decay rate and is assigned to be the length of our clinical trial (90 days). β is the exponential decay rate and will be optimized for each recommendation cycle of each patient using cross-validation.

The above decay mechanisms are both functions of time. However, which function has the better explanation ability of temporal importance needs to be determined via empirical studies. Their comparison and evaluation will be presented at the end of this section.

Also, regularization is employed to address the overfitting problem, possibly caused by the lack of qualified training instances and sparsity (potential $p > N$ problem). Furthermore, the interpretation of the model is also important for practical applications. A simpler model is preferred by the doctors because it can illustrate more clear relationships between the dependent variable and the independent variables. As such, we have adopted a flexible regularization framework called Elastic net [53], which was a compromise between the Ridge-regression (L_2 form) [55] and the Lasso-regression (L_1 form) [56]. The main advantage of the elastic net is that it simultaneously conducts automatic variable selection and continuous shrinkage without losing group effects (variables with high correlations or even identical). For example, carbohydrate intake and total calories are highly correlated. If we employed L_1 penalty, only one feature of the group might be chosen, and the grouped information will be lost thus potentially affect prediction accuracy.

The original cost function $J(\theta)$ is thus extended to a penalized form $J'(\theta)$ as shown in Equation 5.3, along with the regularization penalty parameter λ. Note that P_α is a convex combination of L_1 and L_2 penalty, where α is the elastic net mixing parameter, with $0 \leq \alpha \leq 1$.

$$J'(\theta) = \sum_{i=1}^{N} w_t^i (y_i - \theta_0 - x_i \theta^T)^2 + \lambda P_\alpha(\theta). \tag{5.3}$$

where the elastic penalty $P_\alpha(\theta)$ is defined as:

$$P_\alpha(\theta) = \sum_{j=1}^{p} [\frac{1}{2}(1-\alpha)\theta_j^2 + \alpha|\theta_j|]. \tag{5.4}$$

Cost Function Optimization

The cost function of TWR (Equation 5.3) is optimized using the cyclical coordinate descent, which is an efficient and very fast algorithm to estimate the generalized regression with the convex penalties [54, 57]. Cyclical coordinate descent minimizes $J'(\theta)$ along one gradient direction at a time. The convergence is reached when no improvement is observed after one cycle of line search along all directions.

Here is the detailed optimization process used in TWR. First, all elements of θ are initialized with zero value. Then, suppose we already have the estimates $\widetilde{\theta_0}$ and $\widetilde{\theta_\ell}$ (all $\ell \neq j$), and we would like to partially optimize the cost function $J'(\theta)$ by updating $\theta_j = \widetilde{\theta_j}$ along its gradient direction

$$\frac{\partial J'(\theta)}{\partial \theta j}|_{\theta=\widetilde{\theta}} = -\sum_{i=1}^{N} w_t^i x_{ij}(y_i - \widetilde{\theta_0} - x_i \widetilde{\theta}^T) + \lambda(1-\alpha)\theta j + \lambda\alpha. \tag{5.5}$$

Note that the above derivation only applies to $\theta_j > 0$. There will be a very similar equation for $\theta_j < 0$, which we do not present here.

Then, we can update $\widetilde{\theta_j}$ as follows:

$$\widetilde{\theta_j} \leftarrow \frac{S\left(\sum_{i=1}^{N} w_t^i x_{ij}(y_i - \widetilde{y_i}^j), \lambda\alpha\right)}{\sum_{i=1}^{N} w_t^i x_{ij}^2 + \lambda(1-\alpha)}. \tag{5.6}$$

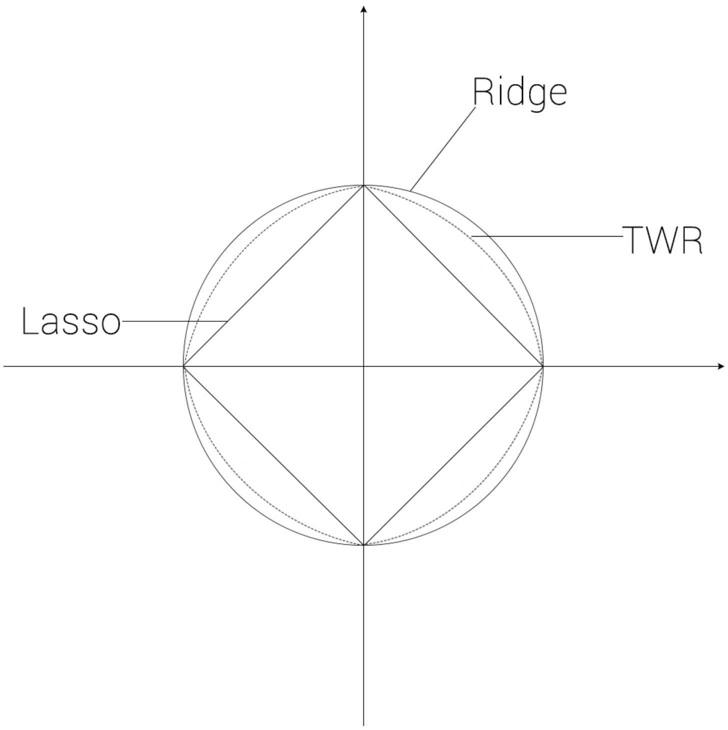


Figure 5.3: Contour plots of Ridge (L2), Lasso (L1), and TWR

where

$$\widetilde{y_i}^j = \widetilde{\theta_0} + \sum_{\text{all}\ell \neq j} x_{i\ell}\widetilde{\theta_\ell}. \tag{5.7}$$

$S(\delta, \gamma)$ is the soft-thresholding operator [58] and calculated as:

$$sign(\delta)(|\delta| - \gamma)_+ = \begin{cases} \delta - \gamma, & \text{if } \delta > 0 \text{ and } \gamma < \delta \\ \delta + \gamma, & \text{if } \delta < 0 \text{ and } \gamma < \delta \\ 0, & \text{if } \gamma \leq |\delta|. \end{cases} \tag{5.8}$$

Such an update process will be repeated in all directions till convergence (the convergence threshold is set to 10E-5).

This cost function has the following properties:

- Convex

  As clearly shown in Figure 5.3, the cost function (Equation 5.3) is strict convex.

- Variable Selection

  The soft-thresholding will give penalties to the coefficient updates and keep the coefficients of irrelevant variables to be zero. As such, the generalized errors could be reduced especially when the training data are insufficient.

- Group Effect

  The group effect means the coefficients of a group of highly correlated variables tend to be equal. It is proved by Hui Zou and Trevor Hastie [53] that $\alpha \neq 1$ (strictly convex) guarantees the grouping effect.

- Temporal Weight

  The designed linear and exponential reweighing mechanisms (Equation 5.2) weigh more on the lifestyle data records closing to recommendation triggered date. Thus, the generated models should reflect more on patients' recent lifestyle activities.

- Computational Efficient

  Each step costs $O(n)$ operations to calculate the gradient. A complete cycle through all p variables costs $O(pn)$ operations. Coordinate descent usually converges within 100 cycles. This fast algorithm is particularly important to generate real-time recommendations to potential large-scale of patients.

Online Parameters Tuning

In practical usage, TWR tunes the proposed two (linear/exponential) weighting mechanisms with different parameter combinations in order to obtain the best model for each recommendation cycle. It assigned the value of α from 0 to 1 with a 0.1 increment and β from 0.5 to 1 with a 0.05 increment to generate a group of parameter combinations. For each combination of α and β, TWR chooses the optimized λ giving the minimum cross-validated error from its candidates sequence. The best model (with optimized α, β, λ) then is chosen to output recommendations.

5.0.4 TWR Model Evaluation

The effectiveness of TWR framework is evaluated mainly via measuring the MAE (Mean Absolute Error) of the blood glucose predictions; other metrics included value range and standard derivation.

We first compare TWR with other state-of-art regression methods including Lasso, Ridge, and Elastic net. They are evaluated using the same patient dataset obtained from our clinical trial (details about this clinical trial will be discussed in the next section). All methods are using the same optimization approach (cyclical coordinate descent). The cost functions of

Table 5.2: Prediction performance comparison of different regression models

Regression Method	MAE
Lasso ($\alpha = 1$)	1.0154
Ridge ($\alpha = 0$)	0.9894
Elastic Net	0.9594
TWR with exponential decay	0.9475
TWR with linear decay	**0.9273**

these methods are shown as follows:

Lasso:

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta_0 - x_i\theta^T)^2 + \lambda\sum_{j=1}^{p}\alpha|\theta_j|.$$

Ridge:

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta_0 - x_i\theta^T)^2 + \lambda\sum_{j=1}^{p}\frac{1}{2}(1-\alpha)\theta_j^2.$$

Elastic Net:

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta_0 - x_i\theta^T)^2 + \lambda\sum_{j=1}^{p}[\frac{1}{2}(1-\alpha)\theta_j^2 + \alpha|\theta_j|].$$

TWR:

$$\sum_{i=1}^{N}w_t^i(y_i - \theta_0 - x_i\theta^T)^2 + \lambda\sum_{j=1}^{p}[\frac{1}{2}(1-\alpha)\theta_j^2 + \alpha|\theta_j|].$$

The results of the comparison are shown in Table 5.2. We can see that by assigning the linear temporal weights, TWR obtains the best accuracy (with least MAE around 0.927) in our experiments. This result is promising if we consider the normal blood glucose range of 4 - 15 mmol/L in our clinical trial. Thus, our results clearly show that TWR outperforms other modern regression approaches for this type of real-world lifestyle dataset. To zoom-in to the detailed blood glucose prediction, we plot the MAE for each patient, shown in Figure 5.4. The x axis represents the patient ID and the y axis represents the MAE. The average value and standard deviations are also shown in Figure 5.4. The best model performance is observed for the patient ID 2 with the MAE 0.67 mmol/L. Even with the worst performance, the MAE is 1.26 mmol/L for the patient ID 7, which is still acceptable, as blood glucose fluctuates throughout the day, and glucometers can have errors just as large. These results are also consistent with

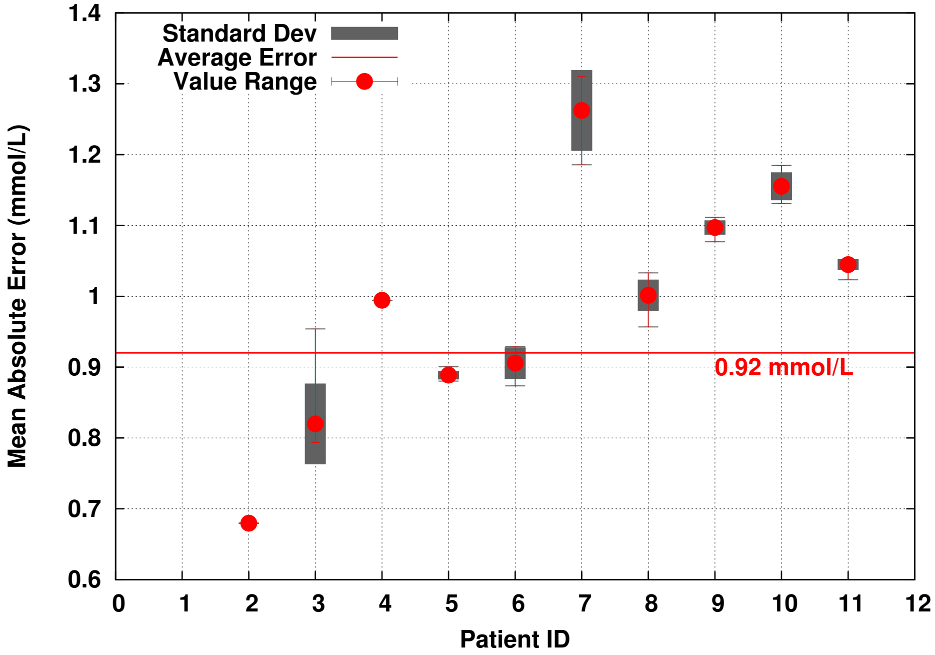


Figure 5.4: MAE of blood glucose prediction for each patient using TWR with linear decay

the state-of-art prediction performance (0.9 $mmol/L$) ([20, 59, 16]) using ANN and expensive CGM data. However, we can achieve similar prediction accuracy using merely lifestyle data and finger-prick blood glucose samples.

Although the MAE is very promising, as discussed above, the R^2 measurement (also called the ratio of deviation explained by the model) during the entire clinical trial is not very high (0.45 ± 0.19). However, predicting human health outcomes is highly challenging and complicated, thus, the results are still satisfiable under the practical circumstance. As we will show in the next section, even such a level of model fitness can still achieve a significant clinical improvement on patients with diabetes.

In summary, the experimental results indicate that our TWR is a reliable and effective framework for analyzing real lifestyle data of diabetic patients. This model, as a important component of GlucoGuide Engine, is validated to effectively predict on patients' blood glucose levels obtained in the clinical trial (see later in Section 5.1).

5.0.5 Model Postprocessing

As we have discussed, for each recommendation cycle, TWR generates a group of candidate models with different parameter settings and choose the best-tuned model with the least MAE to predict the blood glucose levels. With the selected model, it then ranks the features(s) with the largest coefficient (note that all features are already scaled and calibrated) to generate recommendations. Such feature(s) corresponds the maximum impact on the blood glucose levels. Thus, recommendations for changing those feature values can be generated, combined with T2D domain knowledge and patients' historical data.

To illustrate, suppose the discovered linear model is generated as follows:

$$\begin{aligned} BloodGlucose = & \\ & 1.12 \times Carb + -0.88 \times Pro+ \\ & 0.038 \times Fat + 0.056 \times Cals+ \\ & 0.64 \times ProProportion+ \\ & 0.47 \times FatProportion+ \\ & 0.827 \times CarbProportion+ \\ & -0.146 \times StepCount+ \\ & 0.196 \times Aerobic \end{aligned}$$

Clearly, the heath feature with the largest coefficient is the amount of carbohydrate intake. That is, lowering carbohydrate intake should lower the blood glucose levels the most. Thus, the best recommendation is to slowly reduce carbohydrate intake, for this patient. In addition, GlucoGuide will also search our food database to recommend diet replacement based on diet guideline and nutrient equivalence. Such recommendations, after being compared to the diabetes guidelines or verified by our healthcare team, are wrapped in a natural language template and sent to patients' GlucoGuide mobile clients. An example of such a recommendation is shown as follows:

*By examining your uploaded health data we have seen that your **Carbohydrate** and blood glucose levels are highly related. Consider reducing the proportion of carbohydrates you eat at dinner to better control your after dinner blood glucose. For example, try replacing some of the pasta in your meal with an extra portion of lean meat or vegetables, both of which have a lower Carbohydrate content and often help with satiety.*

The complete recommendations list and their validation rules are detailed in Appendix A.

5.1 GlucoGuide Clinical Trial

Preparing and conducting a clinical trial on human subjects is very expensive and resource-intensive. It is estimated that the averaged cost per enrolled subject is slightly more than 6,094 USD (ranged from 2,098 USD to 19,285 USD) [60]. In particular, our clinical trial costs around 4,000 CAD per subject including the cost of technology package (Software development/maintenance, mobile devices with GlucoGuide installed, and medical devices such as glucometers, etc.), laboratory blood tests, training sessions, cellular data communication, subjects follow-ups and questionnaires, etc.

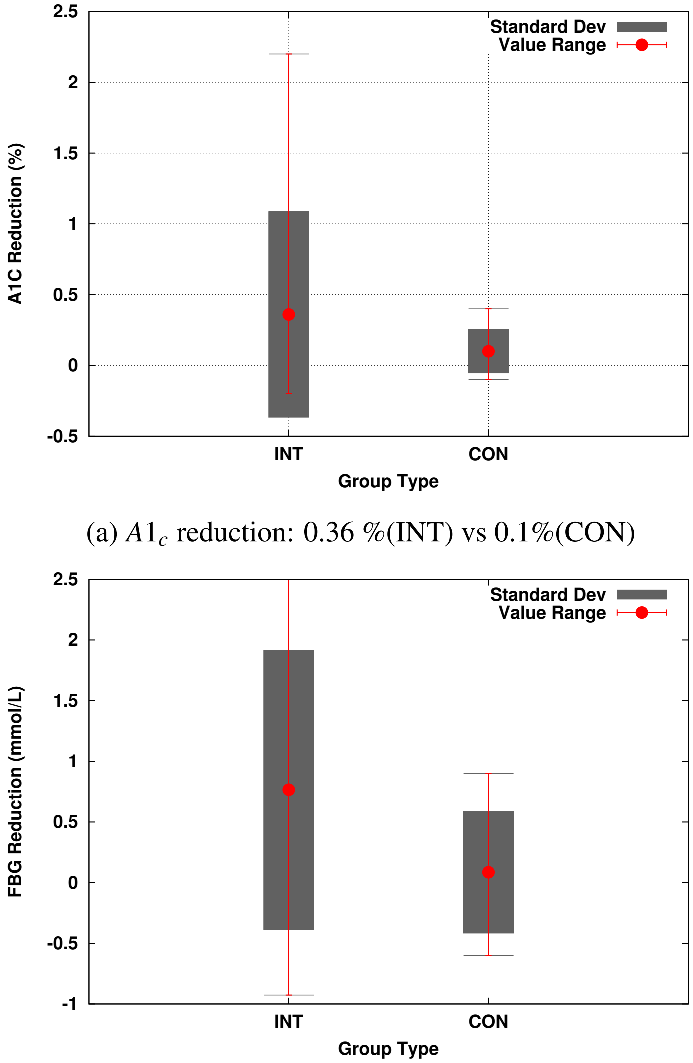
Given the constraint of our budgets, we could only conduct a relatively small clinical trial. However, we set up restrict criteria to ensure the representativeness of our subjects group. The inclusion criteria were: a recent diagnosis of T2D or prediabetes, age between 18 and 80 years, with a survey that concludes a sedentary or low active lifestyle, which was confirmed using the 7-day Physical Activity Recall Questionnaire. In addition, subjects who had difficulties understanding English, were taking more than two diabetes medications, were suffering from severe mental disease or malignant disease, or were abusing drugs were excluded from participation in the study.

At the end of the enrollment, we sampled seventeen adults included seven men (41%) and ten women (59%), with one person in an ethnic minority group (6%) and four individuals with less than a college degree (24%). The average age of the subjects was 62 years (±8). At baseline, men had a mean weight of 105.8 kg (±19.6) and a Body Mass Index [1] of 35.1 kg/m^2 (±6.4). Women had a mean weight of 90.0 kg (±19.9), and a BMI of 34.3 kg/m^2 (±6.3).

Ten subjects are selected to be in the **Intervention Group (INT)**. They are given the GlucoGuide package and taught how to use it in the training sessions. The **Control Group (CON)** includes the rest seven subjects with allied health and physical conditions who do not use GlucoGuide. They are instructed to use standard paper logbooks to keep track of their blood glucose levels, diet, and other similar lifestyle data. Thus, subjects in both groups need to record and keep track of their lifestyle data; the main difference is that the subjects in the intervention group use GlucoGuide with data recording and receive, several times in a week, personalized recommendation while subjects in the control group would not receive any such recommendations. Primary evaluations include two standard clinical blood tests: Fasting blood glucose (FBG) and laboratory-measured $A1_c$ (or $HbA1c$), conducted in the clinical labs (i.e., not at home), for all subjects, before and after the trial. To elaborate, fasting blood glucose is a measure of the amount of glucose in the blood stream after an 8 hour fast, usually overnight. $A1_c$, on the other hand, is proportionally related to the amount of glucose in the blood stream over a long period of time (not affected by day-to-day changes). Therefore, $A1_c$ provides an indication of average blood glucose levels over a longer time, usually three months.

A clinical diagnosis of diabetes is made when fasting blood glucose > 7.0 mmol/L or $A1_c$ > 6.5%, while fasting blood glucose between 6.0 mmol/L and 6.9 mmol/L or an $A1_c$ between

[1] BMI, a measure for human body shape based on an individual's mass and height, T2D patients usually have high BMI



(a) $A1_c$ reduction: 0.36 %(INT) vs 0.1%(CON)

(b) Fasting blood glucose reduction: 0.77 mmol/L (INT) vs 0.086 mmol/L (CON)

Figure 5.5: Comparison of $A1_c$ and fasting blood glucose reduction between two groups, before and after the trial

6.0% and 6.4% is considered prediabetes. A normal person's fasting blood glucose should be less than 6.0 mmol/L, and an $A1_c$ less than 6.0%. (Note that the values of fasting blood glucose and $A1_c$ are not equivalent or equal, but this is beyond the scope of this thesis). By analyzing the differences of $A1_c$ and fasting blood glucose between the two groups, before and after the trial, we can clinically evaluate the effectiveness of GlucoGuide on patients in our clinical trial.

Clearly, for both $A1_c$ and fasting blood glucose values, smaller values are better, and reductions in these values usually represent an improvement in diabetic condition.

Differences in both $A1_c$ and fasting blood glucose reduction between the two groups have been observed before and after the trial. As shown in Figure 5.5a, subjects in the intervention group had an average of 0.36% $A1_c$ reduction compared to only 0.10% in the control group. Such an $A1_c$ reduction for patients with diabetes is clinically significant when considering the differences in $A1_c$ for normal, pre-diabetic, and early diabetic people are quite small. A larger difference can be observed in the fasting blood glucose reduction before and after the trial. For the intervention group, the fasting blood glucose reduction was 0.770 mmol/L, comparing to only 0.086 mmol/L in the control group.

To test the statistical significance, we conducted unpaired T-test on both $A1_c$ and fasting blood glucose reductions between the intervention and control groups. If we use $p = 0.10$ as the significance level, the differences between the two groups on the $A1_c$ reduction and the fasting blood glucose reduction before and after the trial are significant ($p = 0.08$ and $p = 0.06$ respectively). However, our significance levels are just above the stringent level of $p = 0.05$ due to the small sample size. To further evaluate the efficiency of GlucoGuide, a larger clinical trial will be conducted in our future work.

To conclude, patients who use GlucoGuide in our clinical trial reliably improve their diabetic condition after only three months. Roughly speaking, GlucoGuide amounted to turning an early diabetic patient to be pre-diabetic, and pre-diabetic to non-diabetic, in three months.

5.1.1 Adherence to Recommendations

In the best scenario, patients using GlucoGuide should follow or adhere all recommendations made by GlucoGuide. However, most patients in our clinical trial could not adhere 100% to the recommendations made by GlucoGuide. As such, we have defined the Adherence to Recommendations (ATR) ratio for the subjects using GlucoGuide, reflecting how well they adhere the GlucoGuide recommendations throughout the trial. Then we can build an Adherence Model to predict patients' $A1_c$ and fasting blood glucose reduction if they had adhered 100% to the GlucoGuide recommendations.

Here is how we define the ATR (Adherence To Recommendations) ratio. We first divide

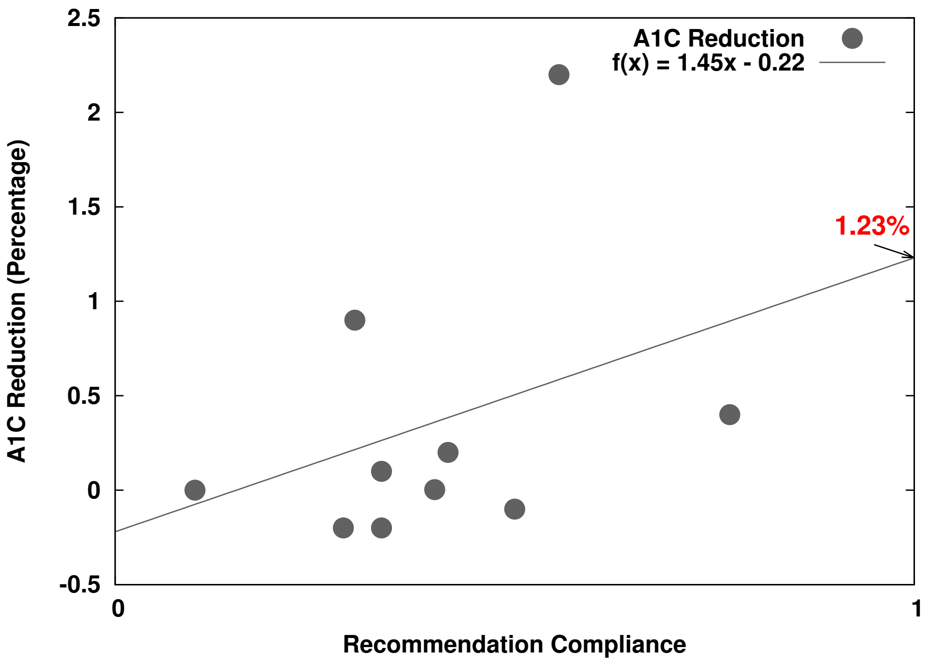


Figure 5.6: Estimated 1.23% $A1_c$ reduction if a patient has 100% ATR

all recommendations into 14 categories. For each category, we design rules to decide if the recommendation is adhered or not. For example, one recommendation type is about reducing the carbohydrate intake. If patients do reduce the carbohydrate in future days, we consider that the patients have adhered to the recommendation. AR is then defined as the ratio on the adhered recommendations over all recommendations sent to the patients. Clearly, the range of ATR is from 0 to 100%, where 100% means that patients adhere completely to GlucoGuide's recommendations.

We build simple linear regression models to predict the $A1_c$ and fasting blood glucose reductions using ATR as an independent variable. Both linear models have positive slopes, indicating that the more the patients adhered to GlucoGuide recommendations, the more $A1_c$ and fasting blood glucose reductions could be achieved. The graph for $A1_c$ is illustrated In Figure 5.6. This Adherence Model predicts that if patients had adhered 100% to GlucoGuide recommendations, the expected reduction on $A1_c$ would be 1.23%. Also, the expected reduction on fasting blood glucose would be 1.03 mmol/L, having patients adhered 100% to GlucoGuide recommendations (graph not shown).

Considering the differences in $A1_c$ and fasting blood glucose for normal, pre-diabetic, and early diabetic people are quite small, people with diabetes would achieve highly significant improvement if they adhere better to GlucoGuide's recommendations.

5.2 Relationships to Previous Works

The use of mathematical models in clinical diabetes is well-known [61], [62]. However, most of the previous research works focus on using predictive models to generate long-term health outcomes such as diagnosis prediction, $A1_c$, complications, etc., as discussed in our literature review. On the contrary, GlucoGuide uses predictive models on short-term health outcomes: lifestyle changes to better control blood glucose levels. Recommending lifestyle changes, as in GlucoGuide, is a form of action mining. It is an important topic in predictive. It aims to generate action plans to maximize the gain or provides personalized recommendations for individuals. Ling et al. [63] first proposed a novel algorithm to suggest actions of changing customers from an undesired status (such as disloyal) to a desired one (such as loyal). However, they used the decision tree model as the target variable is discrete, while in our case, the target variable (the blood glucose levels) is continuous. Thus, an improved regression is used in GlucoGuide.

Qiang et al. [64], [65] extended and refined the decision-tree approach by considering it as a constrained optimization. Their approach is again only applicable to the classification tasks, while GlucoGuide is a regression problem. In addition, those previous works did not consider the involving nature of lifestyle data instances in our work and did not include any temporal information as we did in GlucoGuide.

5.3 Conclusion

The efficiency of self-management for Type-2 Diabetes (T2D) is well-known but remains highly challenging to implement for both patients and doctors in the practice. Our clinical trial shows that GlucoGuide does help T2D patients to alleviate their diabetes conditions based on two standard clinical blood tests. Our adherence model also predicted that the more they adhere to GlucoGuide recommendations, the better the glucose control they would achieve.

In fact, the main difference of diabetes practice between the CON group and INT group is the usage of the entire GlucoGuide system. Thus, it is difficult to point out and compare the clinical efficiency of each sub-system of GlucoGuide (components such as food classification, user interface, Web portal, lifestyle recommendations, etc.) specifically. In the future, more specific clinical trials are needed to distinguish the efficiency of each component of GlucoGuide.

The core function of our lifestyle recommendation framework is the blood glucose prediction. Due to the constrains of resources, the diabetic lifestyle data collected in the clinical trial are limited. It means the TWR framework is probably not converged to his full prediction potentials. In addition, we also want to explore the prediction performance of other non-linear and complex models such as ANN or RandomForest, which usually require a huge amount of training data.

In the next chapter, we will discuss how the small clinical dataset can be expanded and enhanced to large scientifically-controlled datasets for more comprehensive blood glucose prediction tasks.

Chapter 6

Predicting Blood Glucose with Data Expansion

6.1 Blood Glucose Prediction: State-of-the-Art

With nearly two decades' research and development, it is possible now to predict near future blood glucose levels (e.g., 30 or 60 minutes later) within reasonable prediction error ranges. This could be very useful for hyperglycemia/hypoglycemia warnings. For example, if a patient is informed with an extremely high after-meal blood glucose (such as 15 mmol/L), he/she can proactively conduct suitable actions such as increasing moderate/vigorous exercise, and/or adjusting insulin dosage, etc., to stabilize blood glucose beforehand.

The methods of blood glucose prediction reported in the literature can be mainly divided into two groups. The first group uses mathematical models to simulate the physiological dynamics of the glucose-insulin regulatory system. However, due to the complexities of the human body, it is not possible for a simple mathematical model, to precisely predict an individual's blood glucose levels. The second group includes machine learning and data-driven models, which aim to predict blood glucose based on training datasets. The main advantage of data-driven approaches is that they do not require much previous knowledge about the physiology of diabetes. However, a large amount of qualified training data is required for machine learning models to achieve satisfying generalization-errors levels.

Most previous machine learning-based researches focus on continuous blood glucose monitoring (CGM) data. CGM is an advanced technology to measure glucose levels in a real-time throughout the day and night. CGM sensors are inserted under the skin to measure glucose levels in tissue fluid. They are connected to a transmitter sending the blood glucose information via wireless networks. An example of CGM dataset is shown in Figure 6.1, which is a sample of CGM data of a 14 years old T1D patient, sampled every 5 minutes. From Figure 6.1, we can observe a blood glucose peak from 9:50 AM to 11:15 AM after breakfast and be stabilized via insulin injection or medication intake. AutoRegression (AR) [66, 67, 68, 69, 70, 71, 72, 73] is

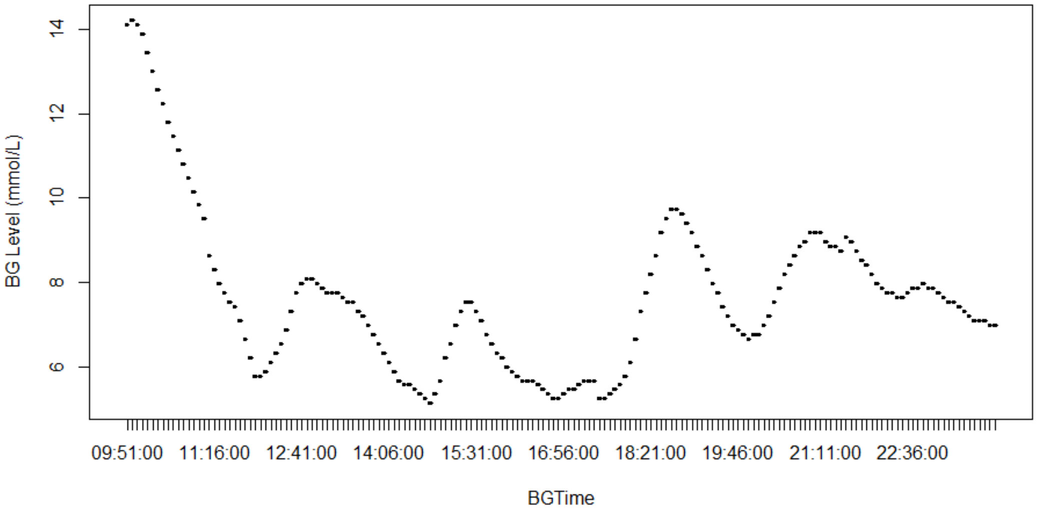


Figure 6.1: CGM data of a 14-years old T1D patient, sampled every 5 minutes

the first proposed and also the most popular framework to predict time-series CGM readings.

A general linear AR model can be formulated as:

$$y(t) = \sum_{i=1}^{M} w_i(t)y(t-i) + e(t). \tag{6.1}$$

where y is the blood glucose at time t, M denotes the order of the AR model, $e(t)$ is as white noise with $E[e(t)] = 0$ and $var[e(t)] = \theta^2$, and w_i is the weight/coefficient of previous blood glucose readings. The future blood glucose prediction $\hat{y}(t+PH)$ by using a weighted combination of history signals before time t:

$$\hat{y}(t+PH) = \sum_{i=1}^{M} w_i(t)y(t-i). \tag{6.2}$$

where PH is the prediction horizon. The coefficient of AR model can be learned usually via least squares-wised algorithms.

To illustrate AR's performance, we used an univariate AR(12) time series model to fit a demo dataset, i.e., the 14-years old T1D patient dataset (around 50,000 readings) by ordinary least squares (OLS) method. The prediction performance is measured by prediction residual.

The mean of prediction residual is 0.11 mmol/L. As we can see from Figure 6.1, the prediction errors ranged from 5 to -5 mmol/L but most are less than 1 mmol/L, indicating AR model is quite effective on this demo dataset.

AR framework has been improving and refining by researchers. Sparacino et al. [67, 68, 70] archived sufficient prediction accuracy by using first-order polynomial AR(1) model, with time-varying parameters (coefficients) learned via recursive least squares (RLS). They also introduced a constant forgetting factor to reduce the weights of blood glucose readings. Eren-Oruklu et al. [71] used AR(3) and auto-regressive with moving average (ARMA) models, with time-varying parameters learned via RLS. Their forgetting factor can be modulated according to the glucose trend. Gani et al. [72] developed an AR(30) model with time-invariant parameters identified by regularized LS. Finan et al. [73] propose an AR model with extra exogenous inputs given by ingested carbohydrates and insulin medications. Artificial Neural Networks (ANN), as a very powerful non-linear predictive framework, has been employing in CGM blood glucose prediction recently [1, 20, 16, 18, 19, 59].

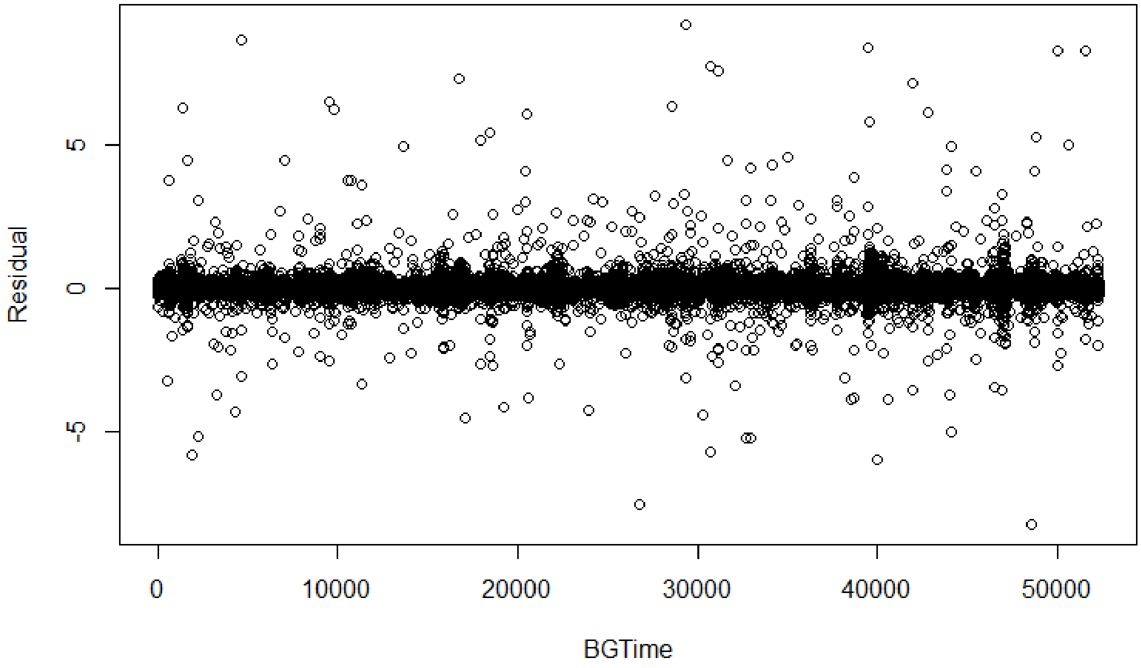


Figure 6.2: AR(12) prediction residual on a 14-years old T1D patient demo dataset prediction

ANN is a modeling tool that consists of simple processing elements, called neurons, linked to each other through weighted connections. The main goal of ANN training is to learn those weights in order to minimize the estimation errors. The main advantage of ANN is that it can learn the relationship between input and output without strong assumptions on the target system. Figure 6.3 shows the basic principle of ANN for CGM blood glucose prediction. Pérez-Gandía et al. [1] first proposed the idea of CGM prediction using ANN in 2010. The inputs of their designed ANN were the values provided by CGM devices during the previous 20 minutes readings, and the outputs were the prediction of blood glucose concentration at chosen PHs (15, 30, and 45 minutes). Their performance was estimated via RMSE (root-mean-square error), with results ranged from 10, 18 and 27 mg/dl (0.55, 1.0, 1.5 mmol/L equivalently) for 15, 30, and 45 minutes of PH respectively. They claimed that ANN model is more accurate than linear AR models.

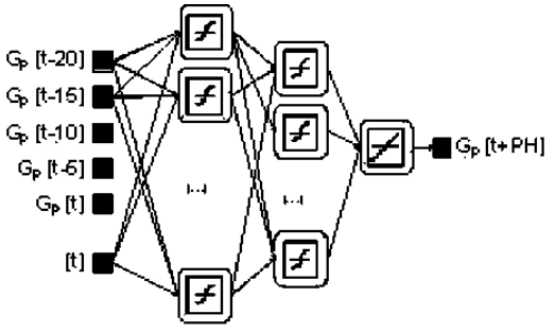


Figure 6.3: ANN architecture for $t+PH$ blood glucose used previous 20 minutes blood glucose readings, cited from the origin paper [1].

Pappada et al. [59] refined ANN-based CGM prediction by adding more inputs such as insulin dosage, nutritional intake, lifestyle, and emotional facts, into ANN model. However, the reported prediction error is relatively high (2.43 mmol/L) due to the imbalance (limited number of hypoglycemic CGM readings) of their training data.

Zecchin et al. [20] proposed a hybrid model combining both an ANN and AR(1) model in parallel to model the nonlinear and linear components of blood glucose. The main rationale is that they believe the performance of the linear AR models decreases for after-meal blood glucose trend (usually blood glucose peaks). Thus, they argue that the relationship between blood glucose and meal information is nonlinear, which is related to the glucose rate of appearance after a meal, modeled by previously published physiological models. As such, they modeled the $\hat{y}(t + PH|t)$, the glucose concentration at $t + PH$ as the sum of two components,

$$\hat{y}(t + PH|t) = \hat{y}_l(t + PH|t) + \hat{e}(t + PH|t). \tag{6.3}$$

where $\hat{y}_l(t + PH|t)$ is the glucose prediction via AR model and $\hat{e}(t + PH|t)$ is the estimation of the error introduced by the AR model, which is modeled via ANN model. Their evaluation results on twenty simulated datasets and nine real datasets showed that the integration of meal information dose improve the prediction accuracy but not significantly.

Later, Zecchin et al. [16] improved ANN-based blood glucose prediction via modifying the architecture of their predictors mentioned in [20]. A modified ANN model called jump NN, i.e., a feed-forward NN whose inputs are not only connected to its next hidden layer, but also to the output layer. Their empirical studies showed that such new model resulted not statistically different than the previous model proposed in [16]. However, they believe that the new jump NN model has much simpler structure.

In addition to AR and ANN framework, the usages of other predictive models such as SVM [74, 69], Regularized Learning [17], and Extreme Learning Machine [69] are also reported in the literature.

In general, we think that the performance of CGM-based blood glucose prediction mainly depends on the quality of training data (volume, sparsity, noise, etc.), the length of PH (typically 30 to 75 minutes), selected predictive framework, parameter tuning, etc. However, the main disadvantage of CGM-based blood glucose prediction is its cost. For example, the CGM datasets used in researches [1, 20, 16, 18, 19, 59, 69] were from a large-scale data-acquisition clinical trial called DIAdvisor project, involving 90 patients across three sites in Europe: France, Italy, and Czech Republic. The first round of this clinical trial has already cost about 7.1 million Euro. Also, they claim that these datasets are the properties of clinical researches and hospitals, meaning they are not publicly accessible. As such, their empirical studies are difficult to be repeated and improved by other researchers.

Moreover, nearly 90% diabetic patients belong to T2D and usually do not need to use CGM devices. The management tools they have nowadays are paper log-book, mobile apps, glucometer, pedometers, insulin pens, etc., which are capable of generating a large scale of non-clinical but diabetic lifestyle data.

6.2 Clinical Data Expansion and Enhancement

It is well known that the main efforts and overhead of conducting machine learning tasks come from data collection and processing. Occasionally, researchers even have to develop a working platform first for collecting training data (such as the GlucoGuide platform). As such, data scarcity problem, also can be considered as “small data“ problem, is one of the main problems of machine learning. The small size of data is often responsible for poor performances and makes the extraction of the significant information for inferences is difficult.

Due to the constraints of resources, the diabetic lifestyle data collected in our clinical trial (detailed in the previous chapter) is not quite comprehensive regarding its dimensions and volume. Fortunately, GlucoGuide, as a commercialized platform now, is collecting lifestyle data accumulatively from the users worldwide. Still, before we have collected sufficient training data, we plan to design data expansion and enhancement mechanisms to create scientifically controlled artificial datasets, as intermediate datasets between the small clinical trial dataset and the large scale of real-world datasets for empirical studies.

In this simulation, we focus on the 2 hours after-meal blood glucose instance because it involves more informative features (such as meals, exercise, etc.). To build the artificial after-meal blood glucose dataset, we first design a new lifestyle feature vector on the basis of features used in the clinical trial (Table 5.1, Chapter 5). We try to use parametric statistical inference to find the properties of underlying distributions and potential dependencies of our clinical dataset. We also apply diabetic domain knowledge to make certain assumptions about the data and determine some important feature dependencies and parameters. Then, we generate unlabeled lifestyle training datasets using the new feature vector. The last step is to send the unlabeled datasets to a popular diabetes simulator called AIDA [75, 76, 77], as an oracle, to generate after-meal blood glucose labels. Since AIDA can process insulin information, we thus can add insulin features, which were not collected in the clinical trial. When the artificial datasets are generated, we can conduct empirical studies to evaluate the quality of the artificial datasets and the prediction performance.

6.3 Artificial Lifestyle Feature Vectors

We categorize the lifestyle health features into three predict levels based on their relevance to blood glucose levels, as shown in Figure 6.4. We believe these predictors are able to cover the main components of patients' daily blood glucose management. Their dependencies and relevance levels in this simulation have been determined via domain knowledge [2]. Note that there are cross-level interactions, and not all predictors are necessarily significant to blood glucose levels. We will explain these predictors and their interactions in the following sections.

6.3.1 L1 Predictions

L1 predictors are patients' general profile information, including:

BMI. Relevant to patients' height information.
Follows continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 35.6$ and $\sigma = 4.3$ inferred from the clinical data sample.

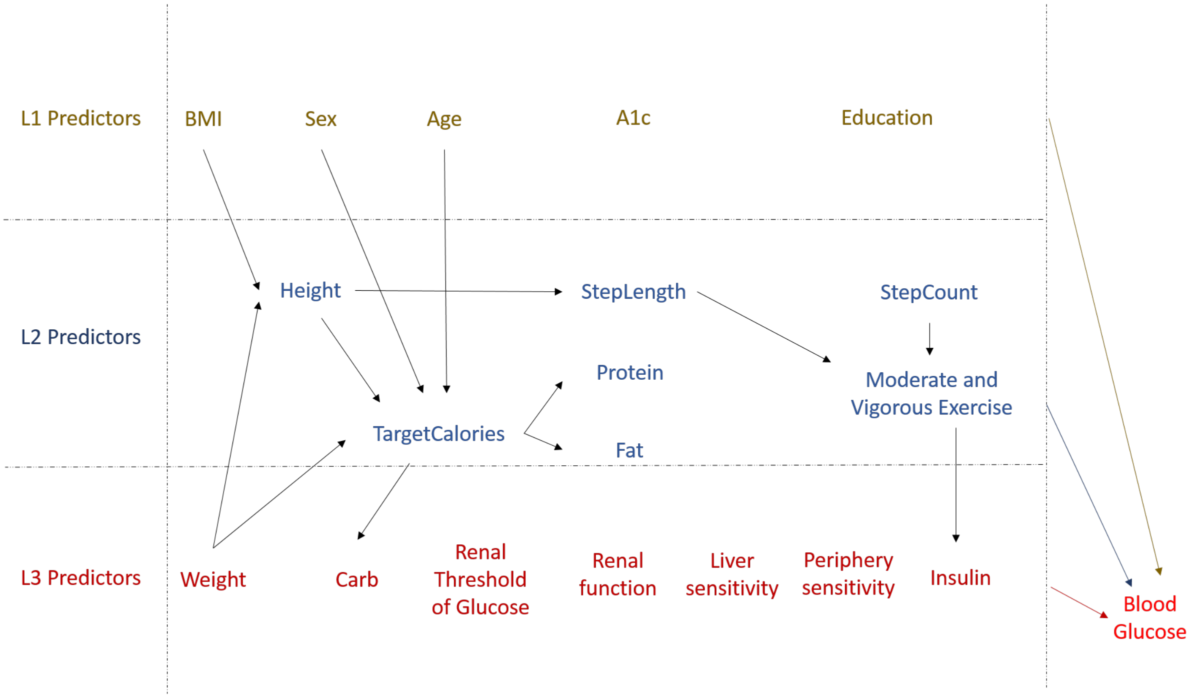


Figure 6.4: The proposed three levels of lifestyle predictors used to predict after-meal blood glucose levels

Sex. Gender information, correlated to daily calories target estimated by Harris-Benedict equation [78], which is a method used to estimate patients' Basal Metabolic Rate (BMR) and daily calories requirements.
Follows discrete uniform distribution $\mathcal{U}$ with 45% to be female and to be 55% male.

Age. Correlated to daily calories target estimated by Harris-Benedict equation.
Follows continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 61$ and $\sigma = 6$ inferred from the clinical data sample.

A1c. Patients' $HbA1c$ information.
Follows continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 0.066$ and $\sigma = 0.009$ inferred from the clinical data sample.

Education. Education level.
Follows discrete uniform distribution $\mathcal{U}$ with 90% to be post-secondary or above and 10% to be below post-secondary.

6.3.2 L2 Predictions

L2 predictors are mostly related to patients' daily activities, containing more predictive information about daily after-meal blood glucose levels, including:

Height. Height information, determined by BMI and weight.

StepLength. Patients' stride length, estimated by Height.

TargetCalories, denoted as C. Energies needed to maintain daily activities, estimated using Weight, Height, Age, and Sex and activities level. In this simulation, daily target calories are distributed equally as 25±5% for breakfast, 25±5% for lunch, 8.75±2.5% for the snack between breakfast and lunch, 8.75±2.5% for the snack between lunch and dinner, 25±5% for dinner, and 8.75±2.5% snack after dinner. It is also designed that patients will have 10% change to skip a meal.

Protein. Daily target protein intake, determined by target calories and calories breakdown ratio. Note that $1g$ protein generates 4 $kcal$ calories.
Follows continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 0.29 * C/4$ and $\sigma = 0.12 * C/4$. The mean and sd value 0.29±0.12 represent the calories ratio contributed by protein, inferred from the clinical data sample.

Fat. Daily target fat intake, determined by target calories and calories breakdown ratio. Note that $1g$ fat generates 9 $kcal$ calories.
Continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 0.21 * C/9$ and $\sigma = 0.18 * C/9$. The mean and sd value 0.21 ± 0.18 represent the calories ratio contributed by protein, inferred from the clinical data sample.

Hourly StepCount. Hourly step count information, related to physical exercise activities. In the previous clinical trial, hourly step counts are collected use Bluetooth-enabled pedometers.

Note that the hourly step counts could have correlations between each other. To test the correlations among step count on different hours, we calculate and plot the covariance matrix in Figure 6.5. As we can see from the Figure 6.5, most pairs have zero or small covariance (less than 0.25). As such, to simplify, we assume the step count on each hour is independent of other hours and we will fit each hour's step count independently. By analyzing the shape and properties (mean, standard deviation, skewness, kurtosis, etc.) of step count samples collected in the clinical trial, we decide to hypothesize three popular non-negative continuous distributions: Weibull $Weillbull(\lambda, k)$, Gamma $\Gamma(k, \theta)$,

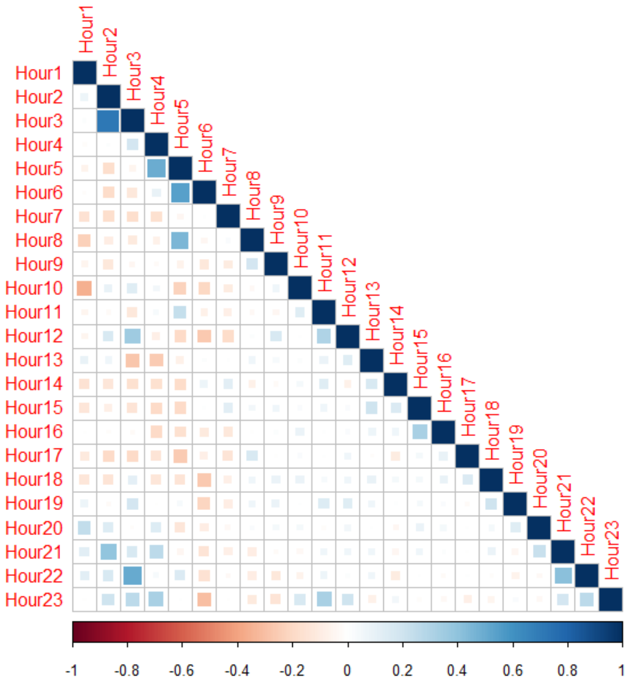


Figure 6.5: The correlation matrix of hourly step counts recorded in the clinical trial

LogNormal $ln\mathcal{N}(\mu,\sigma^2)$ as candidates. Then, we use maximum likelihood estimation to learn their parameters and Aikake's Information Criterion to compare the goodness of fitting. The best fitting distribution found for the step count sample is Gamma distribution. Thus, we believe hourly step count sample

Follows continuous gamma distribution $\Gamma(k,\theta)$ with parameters shape k and scale θ inferred from the clinical data sample (different hour has different parameter pair).

Moderate and Vigorous Exercise. In the previous clinical trial, the only exercise related data collected are step counts from pedometer. However, we also need to determine the specific exercise types and durations. In fact, kinesiology researches show that there is equivalence between a number of steps and physical activities [79]. By referring the step conversion chart proposed by Purdue University [79], we make the following estimations:

4500 hourly steps approximately equivalent to moderate intensity exercise with moderate duration (30 minutes, 150 steps/min).

6500 hourly steps approximately equivalent to vigorous intensity exercise with moderate duration (30 minutes, 216 steps/min).

6.3.3 L3 Predictions

L3 predictors are the most relevant covariates, including:

Weight. Weight information, relevant to BMI.
Follows continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 99$ and $\sigma = 5.0$ inferred from the clinical data sample.

Carbohydrates. Daily target carbohydrates intake, determined by target calories and calories breakdown ratio. Note that 1 g carbs generates 4 $kcal$ calories.
Follows continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu = 0.50 * C/4$ and $\sigma = 0.13 * C/4$. The mean and sd value 0.50 ± 0.13 represent the calories ratio contributed by carbs, inferred from the clinical data sample.

Renal Threshold of Glucose. The renal threshold of glucose (RTG) is the blood glucose concentration at which glucose begins to be excreted by the kidneys into the urine, which is an indicator of kidney function.
Follows discrete uniform distribution $\mathcal{U}$ with 10% to be Low, 80% to be Normal, and 10% to be High.

Renal function. Renal function is a measure of kidney's working status.
Follows discrete uniform distribution $\mathcal{U}$ with 80% to be Normal, 20% to be Reduced.

Liver sensitivity. Liver sensitivity represents the insulin sensitivity of the liver.
Follows discrete uniform distribution $\mathcal{U}$ with 10% to be Normal, 80% to be Reduced, and 10% to be Increased.

Periphery sensitivity. Periphery sensitivity represents the insulin sensitivity of the rest of the body.
Follows discrete uniform distribution $\mathcal{U}$ with 10% to be Normal, 80% to be Reduced, and 10% to be Increased.

Insulin. Patients' daily insulin and dosage information.
Based on the properties of our clinical patients sample, such as aged, overweight, need to reduce food intake, and $A1_c$ 6.6 ± 0.9%, etc. We use the insulin treatment suggested by AIDA as follows:

– *Humulin S, short-acting insulins*. Humulin S is a short-acting human insulin, which usually be taken 20 to 45 minutes before eating. Its peak activity occurs after about 30 minutes and lasts for approximately 2 hours [80]. It is well-known that

Table 6.1: Health features of the 2 hours after-meal blood glucose artificial dataset

Feature Name	Description
MealTime	Timestamp of this meal
MealType	Type of this meal
MealCarb	Carbs intake in this meal
MealCarb Ratio	Percentage of calories from carbs
MealProtein	Protein intake in this meal
MealProtein Ratio	Percentage of calories from protein
MealFat	Fat intake in this meal
MealFat Ratio	Percentage of calories from fat
Calories	Calories of this meal
SnackCarb	Carb intake in the snack between meal and blood glucose testing
SnackProtein	Protein intake in the snack between meal and blood glucose testing
SnackFat	Fat intake in the snack between meal and blood glucose testing
Humulin S Dosage	Active short-acting insulin dosage intake
Humulin I Dosage	Active intermediate-acting insulin dosage intake
Step Count	Step counts between meal and blood glucose testing
Aerobic Step Count	Aerobic step counts between meal and blood glucose testing
Blood Glucose	2 hours after-meal blood glucose level

moderate or vigorous exercise could increase the efficiency of insulin. Thus, we also apply Humulin S adjustment to be 30% reduction for moderate exercise and 50% reduction for vigorous exercise (both for 30 minutes moderate duration). This adjustment mechanism is based on the Guidelines for Insulin Adjustment for Extra Activity suggested by St Joseph's Hospital, London, ON.
Suggested dosage: 4 units before breakfast and 2 units before supper.

– *Humulin I Insulin, intermediate-acting insulins.* Humulin I is an intermediate-acting insulin. It is usually taken in the morning (before breakfast) and/or in the evening (either before dinner or before bed). Its peak activity occurs after about 6 hours and lasts for approximately 10 hours [80].
Suggested dosage: 16 units before breakfast, 12 units before supper, and 14 units before bed.

6.3.4 Assemble Predictors and Construct Feature Vector

Based on the lifestyle predictors explained in the last section, we construct our feature vector of the training dataset for blood glucose prediction, as shown in Table 6.1.

From the above discussion, we can see that the blood glucose prediction technologies allow us to form a training set $z = \{(x_\mu, y_\mu),\ \mu = 1, 2, \ldots, M\}$, $|z| = M$, where

$$\begin{aligned} x_\mu &= ((t_1^\mu, x_1^\mu), \ldots, (t_m^\mu, x_m^\mu)) \ \in \ (\mathbb{R}_+^2)^m\ , \\ y_\mu &= (t_b^\mu, y^\mu) \ \in \ \mathbb{R}_+^2\ . \end{aligned} \tag{6.4}$$

where M is the size of the training datasets and m is the size of the feature vector, t_i^μ is the timestamp of feature x_i in data entry μ, t_b is the timestamp of blood glucose testing, and y^μ is the blood glucose value. Applying the predictive modeling framework, we try to use the lifestyle training dataset z to find the relations between lifestyle data and blood glucose, parametrized with θ

$$f_z(x; \theta) :\ x_\mu \rightarrow y_\mu \tag{6.5}$$

$f_z(x; \theta)$ is constructed as the minimizer of certain regularized or non-regularized cost functions.

6.3.5 Feature Temporal Re-weighting

Since many predictors have time-variant effects (e.g., short-acting insulin reach its peak at 2 hours), we argue that the influence of each predictor on blood glucose level should be adjusted according to their time difference, which is designed as a function of time difference. For each data entry, we scale the decay factor of each feature x_i among all timestamps. Note that this temporal weight is feature-dependent, which is different from the instance temporal weight we designed in the lifestyle recommendation model.

As such, we give any feature x_i with a timestamp associated (e.g., meal time, snack time, etc) an exponential temporal decay $d(x_i)$, to weigh its relevant to the after-meal blood glucose, as shown in Equation 6.6.

$$d(x_i) = \frac{\beta^{(t_b - t_i)}}{\sum_{j=0}^{m} \beta^{(t_b - t_j)}} x_i. \tag{6.6}$$

where x_i is the *ith* feature, t_b is the blood testing time and $\beta = 0.8^{\Delta t}$ is a selected exponential decay function. As such, the Equation 6.5 is rewritten as

$$f_z(x; \theta) :\ d(x_\mu) \rightarrow y_\mu \tag{6.7}$$

In the prediction performance evaluation section, we will show that the temporal feature re-weighting do increase the prediction performance for most popular predictive models.

6.3.6 Blood Glucose Labeling

The target variable is 2-hours after-meal blood glucose level y_μ and labeled via sending lifestyle data entry queries to AIDA simulator. AIDA is a popular diabetes simulator which enables interactive simulation of plasma insulin and blood glucose profiles for demonstration, teaching, self-learning, and research purposes. The AIDA interactive educational diabetes simulator (see `www.2aida.org`) has been proved effective in numerous online-surveys, clinical trials, and empirical studies[75, 76, 81, 82, 77].

AIDA requires all information from L3 predictors to predict 24 hours blood glucose tread, as shown in Figure 6.6. In our case, we only take the discrete 2 hour after-meal blood glucose sample as our labels.

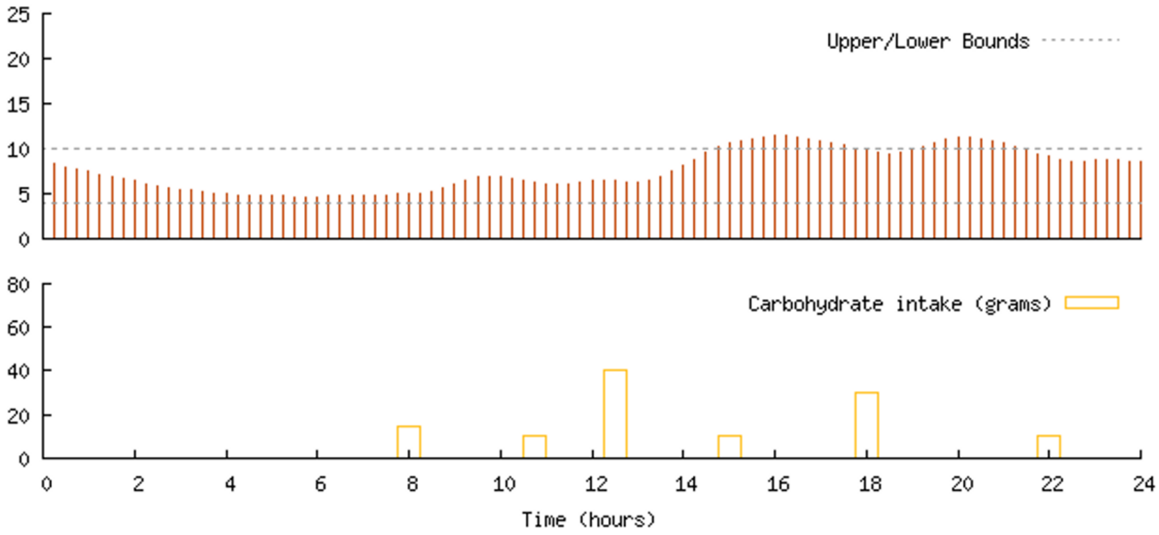
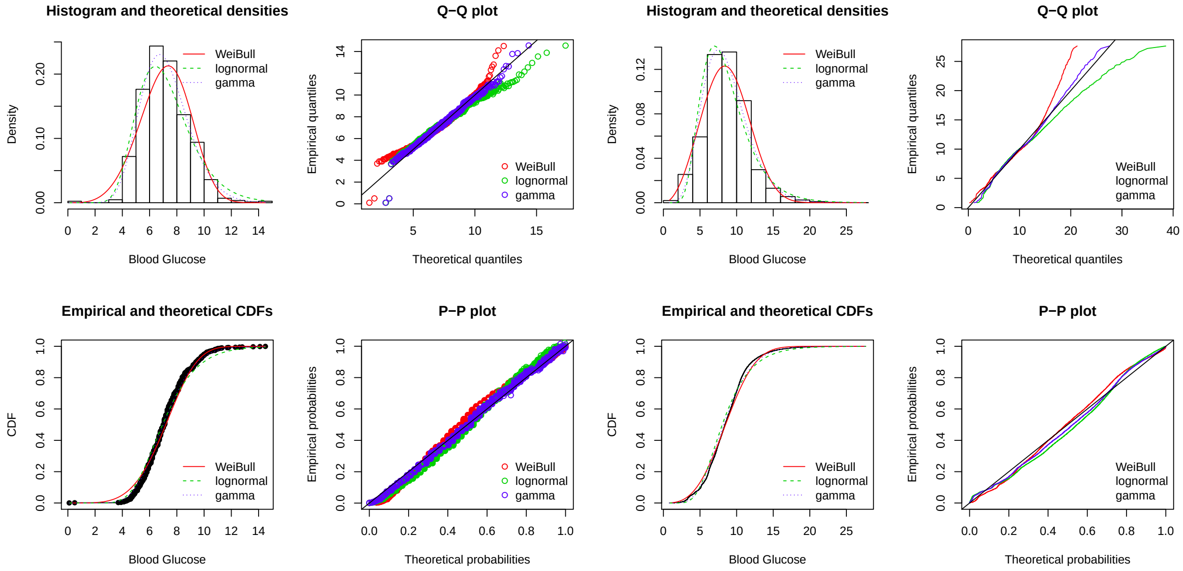


Figure 6.6: 24 hours blood glucose trend, sampled every 15 minutes

6.4 Empirical Evaluation

We evaluate the quality of this data generation process from two aspects. The first one is to compare the statistical properties the blood glucose levels (the target variable) between the clinical samples and artificial samples. Our data generation process is validated if the blood glucose levels in the artificial dataset are statistically similar to clinical trial sample. The second aspect is the predictive capability of artificial datasets. To perform these two evaluation studies, we generated a lifestyle dataset containing 100 patients over three months with 30,000 data entries.

6.4.1 Statistical Properties



(a) Blood glucose levels in the clinical trial sample (b) Blood glucose levels in the artificial dataset

Figure 6.7: Distributions fitting (Weibull, Gamma, and LogNormal) on the blood glucose levels of two datasets (the clinical trial sample and the artificial dataset (100 patients)). Two blood glucose histograms have very similar shapes.

Exploratory blood glucose data analysis is our first step to validate the generated artificial datasets. We use graphical techniques (Histograms/densities, Quantiles-Quantiles plot, Empirical Cumulative Distribution Function (ECDF), and Probability-Probability plots) to visualize the statistical properties of the empirical data and thus hypothesize the candidate families of distributions which would fit well. As we can see in Figure 6.7, blood glucose levels in both dataset are non-negative continuous and nearly normal distributed. Based on these properties, we choose the distribution candidates as $Weillbull(\lambda, k)$, Gamma $\Gamma(k, \theta)$, LogNormal $ln\mathcal{N}(\mu, \sigma^2)$.

The parameters in these candidates are estimated using maximum-likelihood methods. We also use the most popular distribution K-S test to test if the data sample is from a reference distribution statistically. The distribution fitting results show that all candidate distributions have very similar AIC values, but LogNormal distribution has statistical significance in K-S test [1]. The fitted LogNormal distributions and other descriptive percentiles are shown in Table 6.2. As we can see from Table 6.2, the blood glucose levels in the clinical trial sample and

[1] We reject the null hypothesis that the data sample is drawn from the LogNormal distribution if the p-value is less than the significance level (0.05).

Table 6.2: Descriptive percentiles and two fitted LogNormal distributions of blood glucose data in clinical trial and artificial dataset.

Property	Clinical Trial (mmol/L)	Artificial (mmol/L)
Min	0.1	0.82
1st Quantile	6.0	6.6
Median	7.1	8.4
Mean	7.166	8.6
3rd Quantile	8.2	10.2
Max	14.5	27.6
Best-fit	$ln\mathcal{N}(1.94, 0.28^2)$ (K-S test, $p = 0.067$)	$ln\mathcal{N}(2.08, 0.38^2)$ (K-S test, $p = 0.058$)

the 100-patients' artificial dataset share similar statistical properties and distributions, which indicate that our data generation process is valid.

6.4.2 Prediction Capabilities Evaluation

Now the artificial training dataset is well-generated to have similar statistical properties with an original clinical dataset. The next step is to evaluate the predictive ability of this artificial dataset.

In our previous clinical trial, we applied generalized linear-model to predict blood glucose and determine the feature relevance. In this study, since we now have much larger and more informative datasets, we can investigate the prediction abilities of complex and non-linear models. As such, we choose four representative predictive models as: *RandomForest* (discussed in Chapter 4), *ElasticNet* (GLMNET) (discussed in Chapter 5), *ANN*, and *SVM*.

The complete training dataset is randomly split into 5 subsets (10%, 20%, 40%, 60%, and 100%). These percentages are equivalent to the number of patients since the total patient number is 100. These subsets are used to evaluate the scalability of each model. For each sub-dataset, we then split 70% of the data for model training, and 30% for model testing. Also, we have prepared another validation dataset for parameters tuning. The prediction evaluation metric is the standard MAE (Mean Absolute Error). After the standard parameter tuning (grid searching) using the validation dataset, each model is well-calibrated to perform prediction task on the training datasets. We plot their prediction performance in Figure 6.8.

As we can see in the Figure 6.8, the prediction errors of almost all models decrease with the size of the training data. Simple-structured models, such as GLMNET and SVM, seem do

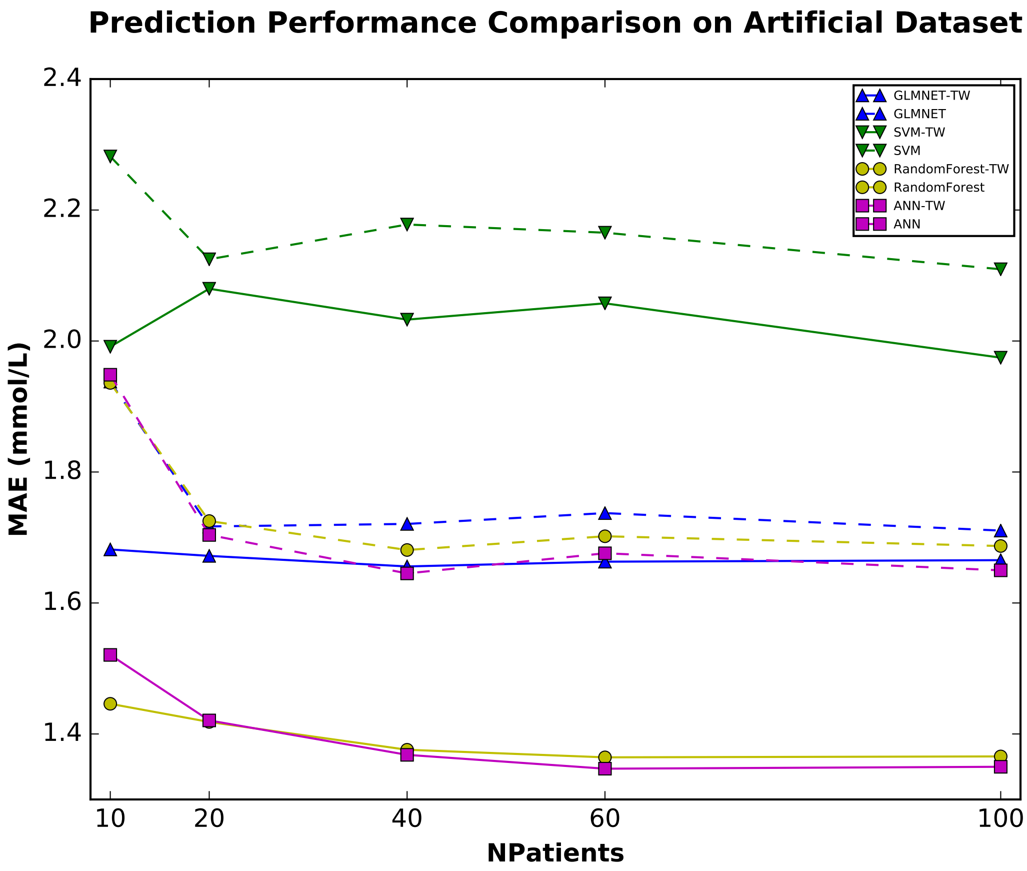


Figure 6.8: Solid lines present the models enhanced with temporal weight (TW), dash lines represent the regular predictive models. We can see that the integration of temporal weights improves the prediction performance for all models. Also, the prediction performance increases with more training data for most models.

not require too much training data to be converged. The start point is ten patients' data over three months, which have the similar size as our entire clinical trial dataset, are sufficient for them to reach their prediction potentials. However, as for complex-structured models such as RandomForest and ANN, they certainly require more data to reach their prediction potentials. Thus, we do believe that our previous clinical dataset might not be sufficient enough to train complex and non-linear models. Inspired by this fact, we argue that deep learning framework such as Convolution Neural Network would achieve better prediction results on a large scale of lifestyle datasets.

We also evaluate the feature temporal re-weighting mechanism. In Figure 6.8, solid lines present the models enhanced with temporal weight (TW) described in Section 6.3.5 and dash lines represent the regular predictive models. Obviously, we can see that the integration of temporal weights does improve the prediction performance of all models. We can see that RandomForest-TW (best prediction error are 1.36571 $mmol/L$) and ANN-TW (1.349877 $mmol/L$) outperform other candidates on this training dataset. These MAE values are larger than the one (0.9273 $mmol/L$) we reported in the clinical trial studies. However, the clinical trial samples in fact have much smaller value range as shown in Table 6.2.

This study also suggests a linear model, even with complex regularization such as ElasticNet, might not be the best prediction model in terms of blood glucose prediction accuracy. However, compared to the complex models such as ANN, the comprehensibility of the linear model is very high. This advantage is in fact important and practically valuable for clinical decision supports.

The artificial blood glucose datasets can also be used for varieties of other machine learning empirical studies such as blood glucose warnings, missing value imputation, data sparsity analysis, transfer learning, etc., as our on-going and future research works. Still, they are not real-world datasets. In order to conduct more reliable empirical studies, we certainly need large datasets with more feature dimensions and huge data volume. Fortunately, the commercialization of GlucoGuide gives us such opportunities, and we can collect large lifestyle data worldwide now. The spin-off of GlucoGuide and our future research objectives will be briefly described in the next chapter.

Chapter 7

GlucoGuide Spin-off

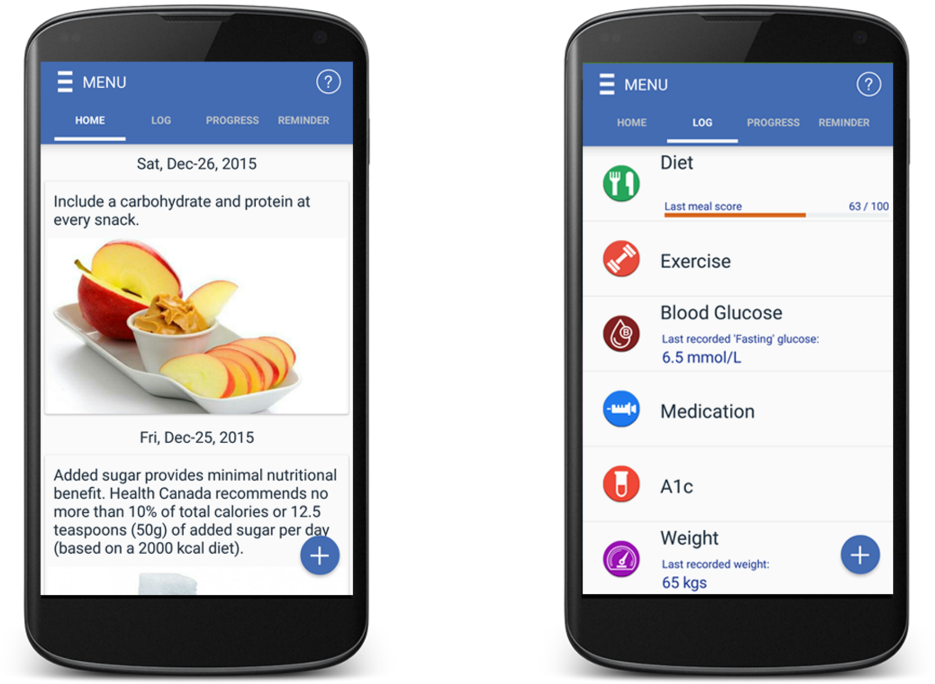
GlucoGuide Corp., established since Feb 2015 (`https://glucoguide.com`), aims to use machine learning technologies to tackle the two main challenges of diabetes self-management, i.e., complexities of data input and lacking of short-term and personalized recommendations, as detailed in Chapter 3. In particular, we have been researching and developing computer vision-based approaches to reduce the complexities of lifestyle data input and provide evidence-based lifestyle recommendations for patients to stabilize their blood glucose effectively.

GlucoGuide has received several rewards and raised approximately 1 million CAD R&D fundings mainly from NSERC I2I, MITACS, and Angel Investments (See Appendix B for details). At this point, the main functionalities of GlucoGuide have been developed, and it is at the stage of promotion and increasing user base. We believe that our user base will explode soon because many GlucoGuide users believe that it is the best diabetes self-management system they have seen. We been collaborating with many health organizations such as the nonprofit organizations like CDA, local Diabetes Education Centers (DECs), and for-profit organizations like MedPoint).

7.1 Current Status

Stable versions of GlucoGuide mobile clients for both Android and iOS can be downloaded from Google Play and App Store respectively and used to log lifestyle data and receive advice and recommendations from data-analytics algorithms. With the current version of GlucoGuide, users are able to:

- Log and track data, including diet, exercise, sleep, blood glucose, insulin, $A1_c$, weight, etc., as shown in Figure 7.1. Users can also use reminder functions to one-click log the data.



Sat, Dec-26, 2015

Include a carbohydrate and protein at every snack.

Fri, Dec-25, 2015

Added sugar provides minimal nutritional benefit. Health Canada recommends no more than 10% of total calories or 12.5 teaspoons (50g) of added sugar per day (based on a 2000 kcal diet).

Last meal score 63 / 100

Last recorded 'Fasting' glucose: 6.5 mmol/L

Last recorded weight: 65 kgs

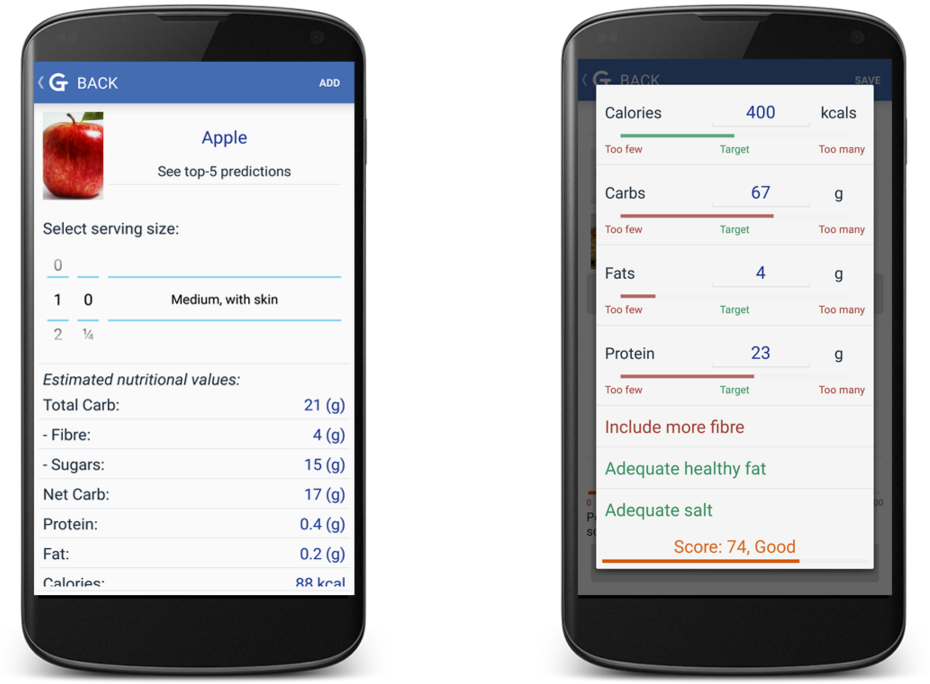
Figure 7.1: Users can use GlucoGuide mobile clients to conveniently log varieties of lifestyle data and receive recommendations

- Receive evidence-based lifestyle recommendations, some examples are also shown in Figure 7.1.

- Log the diet data by snapping a photo of the meal and auto-estimate the nutrition facts, as shown in Figure 7.2;

- Score their meal, based on a novel scoring system, as shown in Figure 7.2. The meal scoring is a novel feature of GlucoGuide system, as it works with the existing data mining system to give individualized feedback based on personal profile (such as weight, height, gender, etc.) Note that the meal scoring system is designed to align with the CDA Guidelines [2] (`http://guidelines.diabetes.ca/Browse/Chapter11`), and with consultation from diabetes specialists and experts.

- View all the lifestyle data entered in a secure online logbook (`https://myaccount.glucoguide.com`). The data are visualized with charts and trends as shown in Figure 7.3, and users can also print their logbook and bring them to their health providers;



Apple

See top-5 predictions

Select serving size:

1 0 Medium, with skin

Estimated nutritional values:

Total Carb:	21 (g)
- Fibre:	4 (g)
- Sugars:	15 (g)
Net Carb:	17 (g)
Protein:	0.4 (g)
Fat:	0.2 (g)
Calories:	88 kcal

Calories	400	kcals
Too few	Target	Too many
Carbs	67	g
Too few	Target	Too many
Fats	4	g
Too few	Target	Too many
Protein	23	g
Too few	Target	Too many

Include more fibre

Adequate healthy fat

Adequate salt

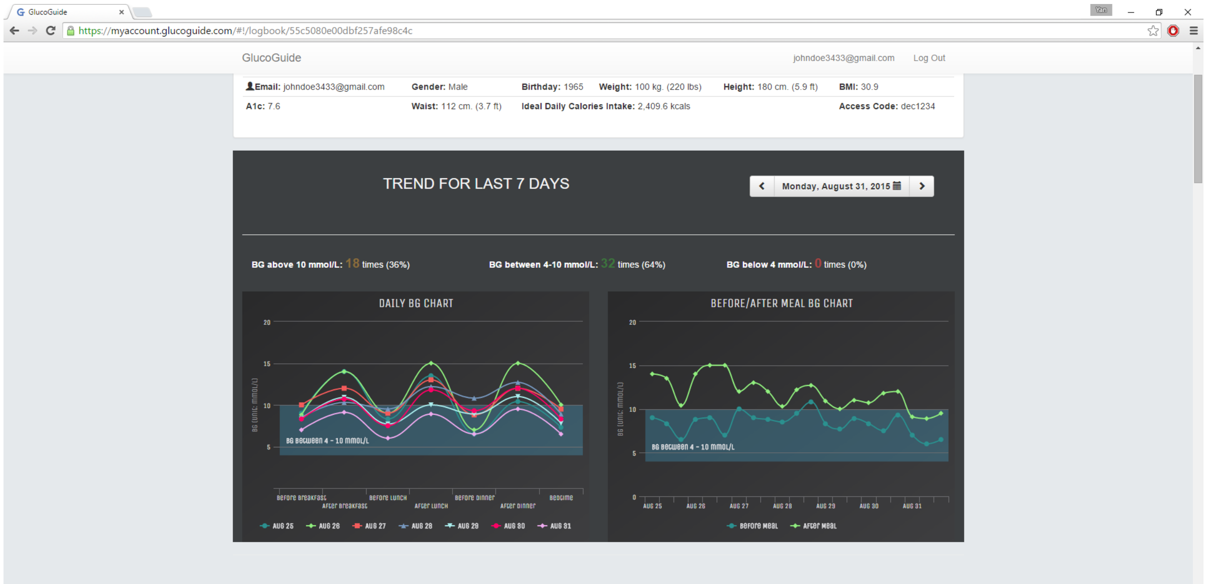
Score: 74, Good

Figure 7.2: Log the diet data by snapping a photo of the meal and auto-estimate the nutrition facts. A overall meal score is also generated to evaluate the whole meal.

7.2 Future Research Objectives

The current GlucoGuide system has already provided a comprehensive platform covering nearly every aspect of diabetes self-management. For example, GlucoGuide is able to record detailed information about each meal including the meal photos, specific food items in the meal, serving and portion size, meal score, etc. As such, our future lifestyle feature vector can be expanded to hundreds of thousands of dimensions. It is possible that in the future we can discover more relationships between food items and blood glucose levels and provide more personalized and detailed recommendations. Such as “Reducing 20% of the portion size of your home-made sandwich to avoid high 2-hour blood glucose level“ or “30 minutes Yoga after dinner would be the most effective way for you to control your fasting blood glucose“.

In fact, for a large feature vector, a huge amount of qualified training data are required to fully describe the varieties of each feature and their relationships. Complex machine learning frameworks are also needed to discover knowledge hidden in the big lifestyle data. Collect such big lifestyle data and mine them with varieties of machine learning frameworks will be our main future research objective.


GlucoGuide
Yan
https://myaccount.glucoguide.com/#!/logbook/55c5080e00dbf257afe98c4c

GlucoGuide johndoe3433@gmail.com Log Out

Email: johndoe3433@gmail.com **Gender:** Male **Birthday:** 1965 **Weight:** 100 kg. (220 lbs) **Height:** 180 cm. (5.9 ft) **BMI:** 30.9

A1c: 7.6 **Waist:** 112 cm. (3.7 ft) **Ideal Daily Calories Intake:** 2,409.6 kcals **Access Code:** dec1234

TREND FOR LAST 7 DAYS

Monday, August 31, 2015

BG above 10 mmol/L: 18 times (36%) **BG between 4-10 mmol/L:** 32 times (64%) **BG below 4 mmol/L:** 0 times (0%)

Figure 7.3: Users can review and print all their data using GlucoGuide online logbook.

Chapter 8

Conclusions

The efficiency of self-management for Type-2 Diabetes (T2D) is well-known but remains highly challenging to implement for both patients and doctors in the practice. Based on the challenges we discovered in the current diabetes management systems, we have proposed three research questions and evaluated a novel system called GlucoGuide to answer them. We summarize the answers to these questions as follows:

- **Can we use classification models to generate real-time food guideline to help patients proactively manage their diet?**
  Answer: Evaluation results of our food classification tool show that it can achieve around 95% classification accuracy using RandomForest model on the proposed feature vector combining textual and nutrient features. Thus, we can conclude that patients could receive empirically reliable real-time food recommendations using our tool. Note that our food classes are subjective and labeled by health experts, which means they are not personalized and be changed over time.

- **Can we predict T2D patients' blood glucose level merely using lifestyle data and discrete fingerstick-based blood glucose samples?**
  Answer: Evaluation results show that the MAEs of the proposed blood glucose prediction framework are similar to the state-of-the-art results on both the real dataset and the artificial datasets, using merely the lifestyle data and discrete blood glucose samples. In fact, to the best of our knowledge, most related blood glucose prediction works are focused on CGM datasets mainly for T1D patients. We could not find the benchmarks of blood glucose prediction for T2D's lifestyle data but we believe there will be more similar research works in machine learning and computational diabetes management areas in the future.

  Also, we find that the prediction performance improved with temporal re-weighting mechanisms. We believe that the performance can be further improved by using deep learning framework on a large scale of lifestyle data.

- **Can we provide clinically effective lifestyle recommendations based on the outcomes of blood glucose prediction?**
  Answer: Our clinical trial results suggest that GlucoGuide system could help T2D patients to alleviate their diabetes conditions based on two standard clinical blood tests. Our adherence model also predicted that the more they adhere to GlucoGuide recommendations, the better the glucose control they would achieve. However, the main difference of diabetes practice between the control group and intervention group is the usage of the entire GlucoGuide system. Thus, it is difficult to point out and compare the clinical effectiveness of each component of GlucoGuide. In the future, more specific clinical trials are needed to distinguish the effectiveness of each component.

Our work can be regarded as a proof of concept in integrating machine learning, mobile computing, and medical knowledge into a mobile intelligent system that can benefit people with chronic diseases, such as diabetes. We hope that our work could inspire future interdisciplinary researchers to apply machine learning and mobile computing into the treatment and management of other diseases.

GlucoGuide now is a university spin-off, allowing us to collect a large scale of practical diabetic lifestyle data in terms of dimensions and volume. We can then design and implement more advanced models to analyze the data and generate more personalized and effective recommendations. We hope our work would have potential impact on the entire diabetes treatment and management area.

Bibliography

[1] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. J. Gómez, M. Rigla, A. DeLeiva, and M. E. Hernando, "Artificial neural network algorithm for on-line glucose prediction from continuous glucose monitoring," *Neural Networks*, vol. 12, no. 1, pp. 81–88, 2010.

[2] "Canadian diabetes association," Canadian Diabetes Association, 2014, http://www.diabetes.ca/.

[3] N. Kaufman, "Information technology in the service of diabetes prevention and treatment use in treatment of diabetes," *International Journal of Clinical Practice*, vol. 65, pp. 47–54, 2011.

[4] R. J. Petrella, C. N. Lattanzio, and T. J. Overend, "Physical activity counseling and prescription among canadian primary care physicians," *Archives of Internal Medicine*, vol. 167, no. 16, pp. 1774–1781, 2010.

[5] C. C. Quinn, J. M. Minor, and A. G. Baldini, "Welldoc mobile diabetes management randomized controlled trial: change in clinical and behavioral outcomes and patient and physician satisfaction," *Diabetes Technology and Therapeutics*, vol. 10, no. 3, pp. 160–168, 2008.

[6] K. Polat, S. Günes, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Systems with Applications*, vol. 34, no. 1, pp. 482–487, 2008.

[7] J. Lindstörm and J. Tuomilehto, "The diabetes risk score: a practical tool to predict type 2 diabetes risk," *Diabetes Care*, vol. 26, no. 3, pp. 725–731, 2003.

[8] K. Chien, T. Cai, H. Hsu, T. Su, W. Chang, M. Chen, Y. Lee, and F. B. Hu, "A prediction model for type 2 diabetes risk among chinese people," *Diabetologia*, vol. 52, no. 3, pp. 443–450, 2009.

[9] J. Park and D. W. Edington, "A sequential neural network model for diabetes prediction," *Artificial Intelligence in Medicine*, vol. 23, no. 3, pp. 277–293, 2001.

[10] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, no. 16, 2010.

[11] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 516–525, 2006.

[12] J. Han, J. C. Rodriguze, and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer," in *Proceedings of the 2nd International Conference on Future Generation Communication and Networking*, 2008, pp. 96–99.

[13] J. Han, J. C. Rodriguez, and M. Beheshti, "Discovering decision tree based diabetes prediction model," *Advances in Software Engineering*, vol. 30, pp. 99–109, 2006.

[14] M. Li and Z.-H. Zhou, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 37, no. 6, pp. 516–525, 2007.

[15] N. H. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114–1120, 2010.

[16] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, "Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information," *Computer methods and programs in biomedicine*, vol. 113, no. 1, pp. 144–152, 2013.

[17] V. Naumova, S. V. Pereverzyev, and S. Sivananthan, "A meta-learning approach to the regularized learningcase study: Blood glucose prediction," *Neural Networks*, vol. 33, pp. 181–193, 2012.

[18] E. Georga, V. Protopappas, A. Guillen, G. Fico, D. Ardigo, M. T. Arredondo, T. P. Exarchos, D. Polyzos, and D. I. Fotiadis, "Data mining for blood glucose prediction and knowledge discovery in diabetic patients: The metabo diabetes modeling and management system," in *Proceedings of 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 5633–5636.

[19] G. Shi, S. Zou, and A. Huang, “Glucose-tracking: A postprandial glucose prediction system for diabetic self-management,” in *Proceedings of 2nd Future Information and Communication Technologies for Ubiquitous HealthCare*, 2015, pp. 1–9.

[20] C. Zecchin, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, “Nerual network incorporating meal information improves accuracy of short-time prediction of glucose concentration,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1550–1559, 2012.

[21] S. Lim, S.-Y. Kim, J. I. Kim, M. K. Kwon, S. J. Min, S. Y. Yoo, S. M. Kang, H. I. Kim, H. S. Jung, K. S. Park, J. O. Ryu, H. Shin, and H. C. Jang, “A survey on ubiquitous healthcare service demand among diabetic patients,” *Diabetes and Metabolism Journal*, vol. 35, no. 1, pp. 50–57, 2011.

[22] L. T. Harris, J. Tufano, T. Le, C. Rees, G. A. Lewis, A. B. Evert, J. Flowers, C. Collins, J. Hoath, I. B. Hirsch, H. I. Goldberg, and J. D. Ralston, “Designing mobile support for glycemic control in patients with diabetes,” *Journal of Biomedicine and Informatics*, vol. 43, no. 5, pp. 37–40, 2010.

[23] S. Krishna and S. A. Boren, “Diabetes self-management care via cell phone: A systematic review,” *Journal of Diabetes Science and Technology*, vol. 2, no. 3, pp. 509 – 517, 2008.

[24] T. Chomutare, L. Fernandez-Luque, E. Arsand, and G. Hartvigsen, “Features of mobile diabetes applications: review of the literature and analysis of current applications compared against evidence-based guidelines,” *Journal of Medical Internet Research*, vol. 13, no. 3, pp. 455–463, 2011.

[25] C. D. Association, “Self-monitoring of blood glucose in people with type 2 diabetes: Canadian diabetes association briefing document for healthcare professionals,” *Canadian Journal of Diabetes*, vol. 35, pp. 317–319, 2011.

[26] J. C. Smith and B. R. Schatz, “Feasibility of mobile phone-based management of chronic illness,” in *Proceedings of 2010 American Medical Informatics Association Annual Symposium*, 2010, pp. 757–761.

[27] P. C. Tang, J. M. Overhage, A. S. Chan, N. Brown, B. Aghighi, M. P. Entwistle, S. L. Hui, S. M. Hyde, L. H. Klieman, C. J. Mitchell, A. J. Perkins, L. S. Oureshi, T. A. Waltimyer, L. J. Winters, and C. Y. Young, “Online disease management of diabetes: Engaging and motivating patients online with enhanced resources-diabetes (empower-d), a randomized

controlled trial," *Journal of American Medical Informatics Association*, vol. 0, pp. 1–9, 2012.

[28] C. C. Tsai, G. Lee, F. Raab, G. J. Norman, T. Sohn, W. G. Griswold, and K. Patrick, "Usability and feasibility of pmeb : A mobile phone application for monitoring real time caloric balance," *Mobile Networks and Applications*, vol. 12, no. 2-3, pp. 173–184.

[29] J.-H. Cho, H.-C. Lee, D.-J. Lim, H.-S. Kwon, and K.-H. Yoon, "Mobile communication using a mobile phone with a glucometer for glucose control in type 2 patients with diabetes: As effective as an internet-based glucose monitoring system," *Journal of Telemedicine and Telecare*, vol. 15, no. 2, pp. 77–82, 2009.

[30] M. Stuckey, E. Russell-Minda, E. Read, C. Munoz, K. Shoemaker, P. Kleinstiver, and R. Petrella, "Data: Results of a remote monitoring intervention for prevention of metabolic syndrome," *Journal of Diabetes Science and Technology*, vol. 5, no. 4, pp. 928 – 935, 2011.

[31] M. Stuckey, R. Fulkerson, E. Read, E. Russell-Minda, C. Munoz, P. Kleinstiver, and R. Petrella, "Remote monitoring technologies for the prevention of metabolic syndrome: the diabetes and technology for increased activity (data) study," *Journal of Diabetes Science and Technology*, vol. 5, no. 4, pp. 936 – 944, 2011.

[32] H.-S. KWON, J.-H. CHO, H.-S. KIM, B.-R. SONG, and S.-H. KO, "Establishment of blood glucose monitoring system using the internet," *Diabetes Care*, vol. 27, no. 2, pp. 478 – 483, 2004.

[33] I. Kouris, S. Mougiakakou, L. Scarnato, D. Iliopoulou, P. Diem, A. Vazeou, and D. Koutsouris, "Mobile phone technologies and advanced data analysis towards the enhancement of diabetes self-management," *International Journal of Electronic Healthcare*, vol. 5, no. 4, pp. 386 – 402, 2011.

[34] D. L. Katz and B. Nordwall, "Novel interactive cell-phone technology for health enhancement," *Journal of Diabetes Science and Technology*, vol. 2, no. 1, pp. 147–153, 2008.

[35] X. Liang, Q. Wang, X. Yang, J. Cao, J. Chen, X. Mo, J. Huang, L. Wang, and D. Gu, "Effect of mobile phone intervention for diabetes on glycaemic control: a meta-analysis," *Diabetic Medicine*, vol. 28, no. 4, pp. 455–463, 2011.

[36] J. hyun Noh, Y. jung Cho, H. woo Nam, J. han Kim, D. jun Kim, H. sook Yoo, Y. woo Kwon, M. hye Woo, J. won Cho, M. hee Hong, J. hwa Yoo, M. jeong Gu, S. ai Kim,

Kyung-eh, S. mi Jang, E. kyung Kim, and H. joon yoo, “Web-based comprehensive information system for self-management of diabetes mellitus,” *Diabetes Technology and Therapeutics*, vol. 12, no. 5, pp. 147–151.

[37] B. Spring, K. Zitouni, D. Harry, N. Moutosammy, A. Sungoor, and B. Tang, “Multiple behavior changes in diet and acitivty a randomized controled trial using mobile technology,” *Archives of Internal Medicine*, vol. 10, pp. 789–796, 2012.

[38] A. Mccallum and K. Nigam, “A comparison of event models for naive bayes text classification,” in *AAAI-98 Workshop on ’Learning for Text Categorization’*, 1998.

[39] L. S. Larkey and W. B. Croft, “Combining classifiers in text categorization,” in *Proceedings of the 19th annual International ACM SIGIR Conference on Research and development in information retrieval*, 1996, pp. 289–297.

[40] V. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[41] N. Gonzalo, “A guided tour to approximate string matching,” *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[42] R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[43] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2002.

[44] C. C. Chang and C. J. Lin, “Libsvm: A library for support vector machines,” 2001.

[45] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] S. B. Kotsiantis, “Artificial intelligence review,” *Journal of Diabetes Science and Technology*, vol. 39, no. 4, pp. 261–283, 2011.

[47] G. R. Institute, “Glycemic index defined,” http://www.glycemic.com/GlycemicIndex-LoadDefined.htm/, 2014, [Online; accessed 24-June-2014].

[48] K. Foster-Powell, S. H. Holt, and J. C. B. Miller, “International table of glycemic index and glycemic load values: 2002,” *The American Journal of Clinical Nutirtion*, vol. 76, no. 1, pp. 5–56, 2002.

[49] L. M. Fischer, L. A. Sutherland, L. A. Kaley, T. A. Fox, C. M. Hasler, J. Nobel, M. A. Kantor, and J. Blumberg, “International table of glycemic index and glycemic load values: 2002,” *American Journal of Health Promotion*, vol. 26, no. 2, pp. 55–63, 2011.

[50] J. Filipovsky, P. Ducimetiere, E. Eschwege, J. L. Richard, G. Rosselin, and J. R. Claude, “The relationship of blood pressure with glucose, insulin, heart rate, free fatty acids and plasma cortisol levels according to degree of obesity in middle-aged men,” *Journal of Hypertension*, vol. 14, no. 2, pp. 229 – 235, 1996.

[51] C. M. Peterson and L. J. Peterson, “Percentage of carbohydrate and glycemic response to breakfast, lunch, and dinner in women with gestational diabetes,” *Journal of Hypertension*, vol. 40 Supplement 2, pp. 172 – 174, 1991.

[52] W. S. Cleveland, S. J. Devlin, and E. Grosse, “Regression by local fitting : Methods, properties, and computational algorithms,” *Journal of Econometrics*, vol. 37, no. 1, pp. 87–114, 1988.

[53] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society*, vol. 67, no. 2, pp. 301–320, 2005.

[54] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2009.

[55] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[56] R. Tibshirani, “Regression shrinkage and selection via lasso,” *Journal of the Royal Society*, vol. 58, no. 1, pp. 267–288, 1996.

[57] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, “Pathwise coordinate optimization,” *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 285–634, 2007.

[58] D. L. Donoho and I. M. Johnstonea, “Adapting to unknown smoothness via wavelet shrinkage,” *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 2009.

[59] S. M. Pappada, B. D. Cameron, P. M. Rosman, R. E. Bourey, T. J. Papadimos, W. Olorunto, and M. J. Borst, “Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes,” *Diabetes Technology and Therapeutics*, vol. 13, no. 2, pp. 135–141, 2011.

[60] E. J. Emanuel, L. E. Schnipper, D. Y. Kamin, J. Levinson, and A. S. Lichter, “The costs of conducting clinical research,” *Journal of Clinical Oncology*, vol. 21, no. 22, pp. 4145 – 4150, 2003.

[61] D. M. Eddy and L. Schlessinger, “Archimedes: A trial-validated model of diabetes,” *Diabetes Care*, vol. 26, no. 11, pp. 3093 – 3101, 2003.

[62] ——, “Validation of the archimedes diabetes model,” *Diabetes Care*, vol. 26, no. 11, pp. 3102 – 3110, 2003.

[63] C. X. Ling, T. Chen, Q. Yang, and J. Cheng, “Mining optimal actions for profitable crm,” in *Proceedings of 2nd IEEE International Conference on Data Mining*, 2002, pp. 767–770.

[64] Q. Yang, J. Yin, C. Ling, and R. Pan, “Extracting actionable knowledge from decision trees,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 43–56, 2007.

[65] Q. Yang, J. Yin, C. X. Ling, and T. Chen, “Postprocessing decision trees to extract actionable knowledge,” in *Proceedings of 3rd IEEE International Conference on Data Mining*, 2003, pp. 685–688.

[66] M. Eren-Oruklu, A. Cinar, D. K. Rollins, and L. Quinn, “Adaptive system identification for estimating future glucose concentrations and hypoglycemia alarms,” *Automatica*, vol. 48, pp. 1892–1897, 2012.

[67] G. Sparacino, F. Zanderigo, A. Maran, and C. Cobelli, “Continuous glucose monitoring and hypo/hyperglycaemia prediction,” *Diabetes Research and Clinical Practice*, vol. 74, no. 2, pp. S160–S163, 2006.

[68] ——, “Continuous glucose monitoring time series and hypo/hyperglycemia prevention: requirements, methods, open problems,” *Current Diabetes Reviews*, vol. 4, no. 3, pp. 181–192, 2008.

[69] Y. Wang, X. Wu, and X. mo, “A novel adaptive-weighted-average framework for blood glucose prediction,” *Diabetes Technology and Therapeutics*, vol. 15, no. 10, pp. 792–801, 2013.

[70] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, “Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 931–937, 2007.

[71] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith, “Estimation of the future glucose concentrations with subject specific recursive linear models,” *Diabetes Technology and Therapeutics*, vol. 11, no. 4, pp. 243 – 253, 2009.

[72] A. Gani, A. Gribok, J. Rajaraman, and J. Reifman, “Predicting subcutaneous glucose concentration in humans: Data-drive glucose modeling,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 246 – 254, 2009.

[73] F. Finan, F. Doyle, C. Palerm, W. Bevier, H. Zisser, L. Jovanovi, and D. Seborg, “Experimental evaluation of a recursive model identification technique for type 1 diabetes,” *IEEE Transactions on Biomedical Engineering*, vol. 5, no. 3, pp. 1192 – 1202, 2009.

[74] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, “A machine learning approach to predicting blood glucose levels for diabetes management,” in *Modern Artificial Intelligence for Health Analytics at Proceedings of 2014 Association for the Advancement of Artificial Intelligence*, 2014.

[75] E. D. Lehmann, C. Tarń, J. Bondia, E. Teufel, and T. Deutsch, “Incorporating a generic model of subcutaneous insulin absorption into the aida v4 diabetes simulator 2. preliminary bench testing,” *Journal of Diabetes Science and Technology*, vol. 1, no. 5, pp. 780–793, 2007.

[76] ——, “Incorporating a generic model of subcutaneous insulin absorption into the aida v4 diabetes simulator 1. a prospective collaborative development plan,” *Journal of Diabetes Science and Technology*, vol. 1, no. 3, pp. 423–435, 2007.

[77] P. Tatti and E. D. Lehmann, “A prospective randomised-controlled pilot study for evaluating the teaching utility of interactive educational diabetes simulators.” *Journal of Diabetes Science and Technology*, vol. 16, no. 1, pp. 7–23, 2003.

[78] A. M. Roza and H. M. Shizgal, “The harris benedict equation reevaluated: resting energy requirements and the body cell mass,” *The American Journal of Clinical Nutrition*, vol. 40, no. 1, pp. 168–182, 2003.

[79] “Convert activity into steps,” http://www.purdue.edu/walktothemoon/activities.html, accessed: 2015-12.

[80] “Humulin insulin guideline,” https://www.lillypro.co.uk/login_form, accessed: 2015-12.

[81] E. D. Lehmann, C. Tarń, J. Bondia, E. Teufel, and T. Deutsch, “Incorporating a generic model of subcutaneous insulin absorption into the aida v4 diabetes simulator. 3. early

plasma insulin determinations." *Journal of Diabetes Science and Technology*, vol. 3, no. 1, pp. 190–201, 2009.

[82] A. Palacio, E. D. Lehmann, and D. E. Olson, "Teaching diabetes to middle-school students with the www.2aida.net aida online diabetes software simulator," *Journal of Diabetes Science and Technology*, vol. 1, no. 1, pp. 106–115, 2007.

Appendix A

T2D Lifestyle Recommendations List

Table A.1: Recommendation List

Type	**Condition**	**Recommendation Templates**
Carbohydrates	Exceeds carbs target and has the largest coefficient	• By examining your uploaded health data we have seen that your carbohydrate and blood glucose levels are related. Consider reducing the proportion of carbohydrates you eat at dinner to better control your after dinner blood glucose. For example, try replacing some of the bread, rice or pasta in a meal with an extra portion of meat or vegetables, both of which have a lower carbohydrate content and often help with satiety. • Blood glucose was a little high after yesterday's meal. Try this trick: Heighten the flavour of your foods with herbs, spices, vinegars, and mustards. They're all low calorie or calorie free and thus don't raise blood glucose levels. • Based on your uploaded data, GlucoGuide has found that carbohydrates are very linked to high after-meal blood sugar. Consider reducing the total carbohydrates of the dinner meal by replacing some of the grains in your meal (ex. rice, pasta, bread) with vegetables or protein.

Protein	Exceeds protein target and has the largest coefficient	• Based on your uploaded data, GlucoGuide has found that your protein intake is linked to your blood glucose levels. To improve your blood glucose levels try eating a smaller portion of protein at meals. Ideally, we should all be eating between 0.8 - 1.8 grams of protein per kilogram of body weight, and protein should make up 20% of our total calories that we eat each day.

Fat	Exceeds fat target and has the largest coefficient	• Blood glucose was a little high yesterday. Think about what you ate: was there a way you could have made it a little healthier? Perhaps scaling back the portion size, decreasing the amount of saturated fat you have at dinner could make the difference. We’ve send a trend between your blood glucose and the amount of fat you eat, so that would be a good place to start! • By examining your uploaded data, we have seen a trend between the amount of fat you have at dinner and your blood glucose levels. Try reducing the portion size of the foods that have a high fat content, and increasing your intake of carbohydrates and protein. For example, in your evening meal, have a salad with garbanzo beans or black beans added in to get extra protein, or try having a 1/2 cup of low-fat or non-fat yogurt with berries for dessert. • To improve your blood glucose levels try eating foods with less saturated fat at meals. Ideally fat should make up 30% of our total calories that we eat each day. Lighten up on fats. For example, decrease the amount of butter, oil, salad dressing, cream cheese, sour cream and other fats you use. They’re loaded with calories and some have unhealthy saturated fat.

Carb Ratio	Carb ratio exceeds target ratio and has the largest coefficient	• Great healthy eating choices - by incorporating whole grains and protein-rich Greek yogurt you are fueling your body with the nutrients you need. Keep seeking out ways to add those vegetables for extra punch.
Protein Ratio	Exceeds protein ratio target and has the largest coefficient	• Based on your uploaded data, GlucoGuide has found that your protein intake is linked to your blood glucose levels. To improve your blood glucose levels try eating a smaller portion of protein at meals. Ideally, we should all be eating between 0.8 - 1.8 grams of protein per kilogram of body weight, and protein should make up 20% of our total calories that we eat each day.
Fat Ratio	Exceeds fat ratio target and has the largest coefficient	• Great work incorporating an extra serving of veggies to your meals. Looking to lighten up on saturated fats. Look at the nutrition facts of butter, oil, salad dressing, cream cheese or sour cream. Some varieties are loaded with calories and too much unhealthy saturated fat. Consider reducing your portions or replacing them with lower-fat varieties.

Step Counts	Detected inactive and has the largest coefficient	• Your blood glucose levels have been a little high before dinner the last two days. Consider adding a short walk before dinner or between dinner and your blood glucose measurement two hours after dinner.
Aerobic	Detected inactive and has the largest coefficient	• Your blood glucose levels have been a little high before dinner the last two days. Consider adding a moderate exercises before dinner or between dinner and your blood glucose measurement two hours after dinner. • Physical activity is anything that increases your heart rate. It is beneficial as it improves blood flow to organs and helps body in the production of insulin. Try reaching your training heart rate today to improve your cardiovascular health!
Blood Glucose Reminder	Missing blood glucose testing for adjacent three days	• Please remember to upload your evening blood glucose: before and two hours after dinner.

Data Record Reminder	Missing data uploading for adjacent three days	• Please remember to record and upload your evening data: blood glucose before and two hours after dinner, what you ate, your blood pressure and step count. • We've missed you the past couple days. Please upload your dinner, blood glucose, blood pressure and step count today. Way to meet your step count goals the last few days!

Blood Glucose Target	Reach blood glucose targets	• Yesterday's blood glucose levels look great. Way to go! Healthy choices, incorporating lots of veggies and fruits means that your body will get the nutrients it needs. • Awesome - blood glucose levels are looking really good. • Please upload what you've been eating for dinner. Blood glucose is looking great lately, and we want to see how you've achieved such success! • You are doing well - blood glucose and blood pressure are looking good today. Based on your uploaded data, GlucoGuide has found that carbohydrates are very linked to high after-meal blood sugar. Continue to replace some of the grains in your meal (ex. rice, pasta, bread) with vegetables or protein. • Your blood glucose levels have been a little more elevated than normal the last few days. Is there something you're doing differently? • Your blood glucose was very high this morning after your meal. This may happen once in a while, but it is better to keep levels steady. Food that has fibre, protein, and healthy fats helps us do that. For example, eating a meal with veggies and hummus instead of chips and dip will keep your blood glucose levels lower. • Your blood glucose was quite low - if you aren't planning on having dinner, have a small snack, with about 15 g of carbohydrates, such as an apple, an orange or a pear.

Blood Glucose Warning	Abnormal blood glucose levels detected	• It looks like your blood glucose was very variable today. To even out blood glucose swings, try to eat smaller meals every 2-4 hours that include some protein, fat, carbohydrates and fibre. • It looks like your blood glucose was very variable today. To even out blood glucose swings, try to eat smaller meals every 2-4 hours that include some protein, fat, carbohydrates and fibre. • Your fasted blood glucose was quite low tonight - we would like to monitor this. Please take your blood glucose two hours after your meal. • Your after dinner glucose level was very low tonight. We'd like to monitor this. Please have something to eat with at least 15 grams of carbohydrates and repeat the measure in two hours. • You are making great healthy choices! Remember to reward yourself - decide what kind of reward would work best for you: maybe it's praise from your doctor, a new pair of running shoes, or some time for reading a good book. Use the reward you choose to treat yourself when you've reached a goal.

Step Count Target	Reach step count target	• Great work uploading your data. Your daily step counts are looking good: small steps towards change, added together can yield big results for your health and wellness. • Blood pressure is looking good! Keep it up with that step count, yesterday - 8000+! Fantastic. • Small changes today will yield great results when added together. Your step count is great, the next step is adding those veggies and healthy snacks! • Go for that step goal! Plan ahead: set aside time in your day for activity, and you're more likely to follow through with it. • 10,000+ steps yesterday! Awesome! • Blood glucose levels and step count are looking great. Keep it up - your healthy lifestyle can help reduce your risk of disease!

Encouragement	No data driven recommendation generated and/or overall well performed	• Make small goals and then plan to celebrate when you accomplish them! Reward is an important part of goal setting. Ideas include: a relaxing walk, a movie, a good book, or a chat with an old friend. • Good work uploading your data. By uploading accurate and complete information about your dinners we will be able to see what affects your blood glucose most. • Fantastic job uploading your data. Through healthy eating and physical activity you are directly and positively affecting your health. Remember - slow and steady wins the race! • BP looked good yesterday! You can now move to just monitoring dinner. Keep on uploading your dinner data accurately and completely, so that we can find what is affecting your blood glucose most.

Appendix B

Rewards and Fundings

- The 1st prize for the best Clinical Research Presentation at the 2nd Annual Diabetes Research Day by the Schulich School of Medicine & Dentistry in Nov 2011.
- 79,000 CAD R&D funding from 2014 NSERC Idea to Innovation Grants.
- Angel Investments from Jordann Capital Management Inc since 2014.
- 125,000 CAD R&D funding from 2016 NSERC Idea to Innovation Grants.
- 146,000 CAD R&D funding from 2016 MITACS Accelerate.

Curriculum Vitae

Name:	Yan Luo
Post-Secondary Education and Degrees:	University of Western Ontario 2011 - 2015 Ph.D. Candidate Memorial University of Newfoundland 2009 - 2011 M.Sc Nanjing University of Aeronautics and Astronautics 2004 - 2008 B.Sc
Related Work Experience:	Teaching and Research Assistant The University of Western Ontario 2011 - 2015 Core Researcher and Developer GlucoGuide Corp. 2014 - 2015 Teaching and Research Assistant Memorial University of Newfoundland 2009 - 2011 Software Engineer Pingan Bank 2008 - 2009

Publications:

- Luo, Yan and Ling, Charles. GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing. 2014 IEEE International Conference on Data Mining.

- Luo, Yan and Ling, Charles and Ao, Shuang. Mobile-based Food Classification For Type-2 Diabetes Using Nutrient and Textual Features. 2014 IEEE International Conference on Data Science and Advanced Analytics.
- Petrella, Robert J.; Schuurman, Jody; Luo, Yan; Ling, Charles X (2014). A Smartphone-based Personalized System for Alleviating Type-2 Diabetes. American Telemedicine Association 2014.