

# Glucoguide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing

Yan Luo	Charles Ling, Ph.D.	Jody Schuurman	Robert Petrella, MD, Ph.D.
Computer Science Dept. Western University London, Canada yluo84@csd.uwo.ca	Computer Science Dept. Western University London, Canada cling@csd.uwo.ca	School of Kinesiology Western University London, Canada jodyschuurman@gmail.com	Schulich School of Medicine and Dentistry Western University London, Canada petrella@uwo.ca

**Abstract**—Type-2 Diabetes (T2D) is a dreadful disease affecting hundreds of millions of people worldwide, and is linked and worsen by unhealthy lifestyles. However, managing T2D effectively with lifestyle change remains highly challenging for both T2D patients and doctors. In this paper, we proposed, built, and evaluated a personalized diabetes recommendation system, called Glucoguide for T2D patients. Glucoguide conveniently aggregates a variety of lifestyle data via medical sensors and mobile devices, mines the data with a novel data-mining framework, and outputs personalized and timely recommendations to patients aimed to control their blood glucose level. To evaluate its clinical efficiency, 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. Two standard laboratory blood tests for diabetes were 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.

## I. INTRODUCTION

Diabetes is a metabolic disease in which patients have abnormally high blood glucose (BG). 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, or the human body does not properly use the insulin it makes [1], [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.

After a person is diagnosed with T2D, usually medications will be prescribed by doctors. In addition, a protocol called Self-Monitoring of Blood Glucose (SMBG) will be advised by healthcare providers to help the patient achieve proper BG targets [1]. In fact, BG is affected by a large number of factors, including what you ate, how long ago you ate, your starting blood glucose level, physical activity, mental

stress, illness, sleep patterns, and so on [1]. Each factor has different impact on a specific person's BG. SMBG requires T2D patients to control these factors including check their BG regularly, restrict and accurately count the amount of carbohydrates, protein, fat, and so on, in their diet, along with other types of lifestyle changes. Note that SMBG is designed for general T2D patients, thus not personalized. Often after the initial meetings with the diabetic specialists and dieticians, the T2D patients are on their own in applying SMBG on the daily basis.

Adopting and maintaining healthy lifestyle changes is highly challenging for Type-2 diabetes patients [3]. For example, the overwhelming complexity of carbohydrate or calories counting presents an often insurmountable obstacle for most T2D patients. Also, when the patient's BG is high, it is usually 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 in recent years has provided patients with more convenient data recording and management tools, as well as possible real-time support from their healthcare providers. However, there are still large hurdles preventing the current mobile-based diabetes management systems from reaching their full potentials. Most smartphone softwares (apps) in the App Store and Google Play Store, such as Glooko™ and Glucose Buddy™, only record and plot blood glucose and life-style data, without data analytics on patient data to actively advise and engage patients. A few others, such as WellDoc [4], do give patients personalized advice but only prescribed by doctors who need to access the system to review their patients data. There is still no or little data mining capability in these systems.

In this paper, we present Glucoguide, a data-driven, data-mining based lifestyle management solution for T2D. Glucoguide conveniently collects a variety of lifestyle data (diet, exercise, etc.) via medical sensors and wearable devices, and uploads the data securely to our computing server. If the data indicate that a patient is in emergent or dangerous situations, such as abnormal high or low BG levels, Glucoguide provides immediate assistance to the patient, as well as alerts to our healthcare team.

If patients' data do not present emergent or dangerous situations as above, data will be collected and uploaded to our computing server. A few times in a week, a novel data-mining framework called temporal-weighted regression (TWR) will be deployed in the server to discover correlations between the recent lifestyle data and the BG levels for each patient. Such correlations will be framed in natural language templates, and sent to patients' smartphones as recommendations. As patients' data are different, the recommendations are also personalized.

To evaluate the performance of TWR and the clinical efficiency of GlucoGuide, 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. Using the data collected during the clinical trial, we empirically evaluated the TWR framework and showed that it outperformed other state-of-the-art regression models. 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 amounted to turning an early diabetic patient to be pre-diabetic, and pre-diabetic to non-diabetic, after a three-month trial. See Section III for details.

The rest of the paper is organized as follows: Section II presents our the design principle of GlucoGuide solution, Section III describes the clinical trial. The last two sections discuss the relationships to other works and conclusions.

## II. GLUCOGUIDE SOLUTION

The GlucoGuide system contains three main components, including the 1) Bluetooth-enabled medical devices, 2) GlucoGuide smart phone application (App, available in Google Play), and 3) GlucoGuide Engine (Computing Server), which deploys the data mining framework.

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

Since patients uploaded various types of data simultaneously or at any times of a day, some of these data were very noisy and abundant. As such, the data aggregation and preprocessing module was designed to convert the raw lifestyle data into proper data formats for further data mining tasks. Then, the proposed TWR (Temporal-Weighted Regression) was able to identify the temporal patterns and convert them into computer-generated recommendations.

### A. Temporal-Weighted Regression (TWR)

Designing and adopting an efficient, robust, and reliable model to mine real T2D patients' data steam is challenging. Firstly, T2D lifestyle data is costly since they are from a variety of sensors, food database queries, real BG samples from patients, which indicates that the scale of such dataset could not be very 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 uploading. This causes 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 (BG) and independent variables were all numeric, regression analysis thus became the most suitable solution for our problem domain at this point. In addition, regression with regularization is well-known to prevent model overfitting, and generate models for better interpretation.

Although regression with regularization has been researched for years [5]–[10], we still make the following novel 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 it carries. 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 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 could also inspire other researchers to design better data mining applications in other telemedical domains.

At a high level, TWR was deployed in the GlucoGuide Engine and triggered once every few days to generate lifestyle recommendations for each patient. In every recommendation cycle, it generated 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 chose the most relevant health feature(s) and convert it into recommendations using predefined NLP templates.

TWR is 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 [5]. 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; those far away receive low ones.

In practical usage, TWR tuned two (linear/exponential) temporal weighting mechanisms with different parameter combinations in order to obtain the best model for each recommendation cycle.

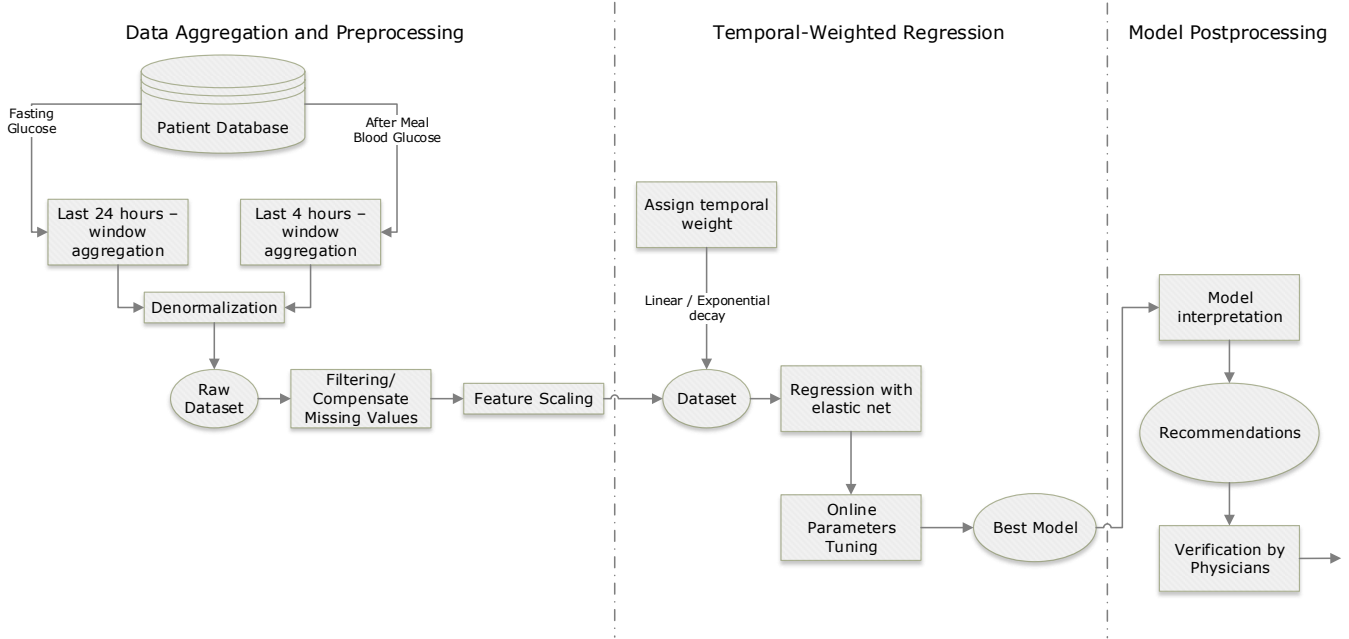


Fig. 1: T2D lifestyle recommendations framework

### B. TWR Model Evaluation

The accuracy and efficiency of TWR framework were evaluated mainly via measuring the MAE (Mean Absolute Error) of the BG predictions; other metrics included value range and standard derivation.

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

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

#### TWR:

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

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

TABLE I: Comparison of different regression models. Note that  $\alpha$ ,  $\beta$ , and  $\lambda$  usually have different values for each recommendation cycle of each patient, thus we do not list them all here

The results of the comparison are shown in Table I. 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 usual BG range of 4 - 13 (mmol/L). This indicates an estimated prediction accuracy of 89.7%. Thus, our results clearly show that TWR outperforms other modern regression approaches for this type of real-world patient datasets.

To zoom-in to the detailed BG prediction, we plotted the MAE for each patient, shown in Figure 2. 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 2. 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 BG fluctuates throughout the day, and glucometers can have errors just as large.

Although the MAE is very promising, as discussed above, the  $R^2$  measurement during the entire clinical trial is not

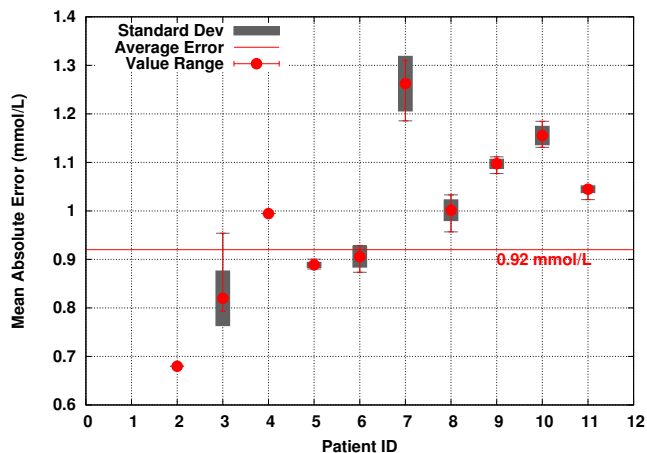


Fig. 2: MAE of BG prediction for each patient using TWR with linear decay

very high ( $0.45 \pm 0.19$ ), suggesting linear models may be too simple. 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 modest model fitness can still achieve a significant clinical improvement on patients with diabetes.

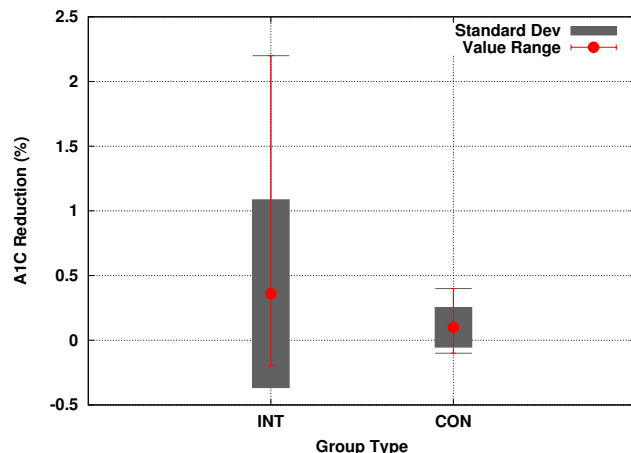
In summary, the experimental results indicate that our TWR is a reliable and efficient framework for mining real-life data of diabetic patients. This model, as the core of GlucoGuide system, is validated to effectively predict on patients' BG levels obtained in the clinical trail (see later in Section III).

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

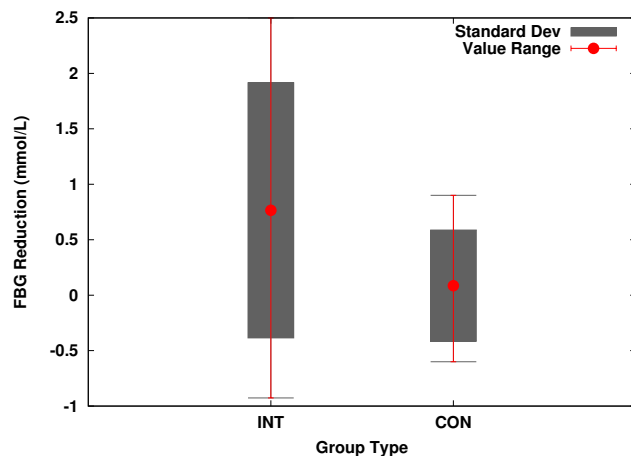
### III. CLINICAL TRIAL

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) [11]. In particular, our clinical trial costs around 3,375 USD per subject including the cost of technology package (Software development/maintenance, Android Smartphone with GlucoGuide installed, and medical devices such as glucometers, etc), laboratory blood tests, training sessions, cellular data communication, subjects follow-ups and questionnaires, and so on.

Given the constraint of our budgets, we could only conduct a relatively small clinical trial. However, we set up a restrict criteria to ensure the representativeness of our subjects group.



(a)  $A1c$  reduction between two groups: 0.36% (INT) vs 0.1% (CON)



(b) FBG reduction between two groups: 0.770 mmol/L (INT) vs 0.086 mmol/L (CON)

Fig. 3: Comparison of  $A1c$  and FBG reduction, before and after the trial

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 2 diabetes medications, were suffering from severe mental disease or malignant disease, or were abusing drugs were excluded from participation in the study.

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 would receive, several times in a week, personalized recommendation from our data-mining algorithms described earlier, while subjects in the control group would not receive any such recommendations.

Primary evaluations included two standard clinical blood tests: Fasting BG (FBG) and laboratory-measured  $A1c$  (or  $HbA1c$ ), conducted in the clinical labs, for all subjects, before and after the trial. To elaborate, FBG 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 in 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  $FBG > 7.0$  mmol/L or  $A1_c > 6.5\%$ , while  $FBG$  between 6.0 mmol/L and 6.9 mmol/L or an  $A1_c$  between 6.0% and 6.4% is considered prediabetes. A normal person's  $FBG$  should be less than 6.0 mmol/L, and an  $A1_c$  less than 6.0%. By analyzing the differences of  $A1_c$  and  $FBG$  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  $FBG$  values, smaller values are better, and reductions in these values usually represent an improvement in diabetic condition. Differences in both  $A1_c$  and  $FBG$  reduction between the two groups have been observed before and after the trial. As shown in Figure 3a, subjects in the intervention group had an average of 0.36%  $A1_c$  reduction compared to only 0.10% in the control group ( $p = 0.08$ ). 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  $FBG$  reduction before and after the trial. For the the intervention group, the  $FBG$  reduction was 0.770 mmol/L, comparing to only 0.086 mmol/L in the control group ( $p = 0.06$ ).

To conclude, participants who used GlucoGuide in our clinical trial reliably improved 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.

#### IV. RELATIONSHIPS TO PREVIOUS WORKS

Recommending lifestyle changes, as in GlucoGuide, is a form of action mining. It is an important topic in data mining. It aims to generate action plans to maximize the gain or provides personalized recommendations for individuals. Ling et al. [12] 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. [13], [14] 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 data instances in our work, and did not include any temporal information as we did in GlucoGuide.

#### V. DISCUSSIONS AND CONCLUSIONS

The efficiency of self-management for Type-2 Diabetes (T2D) is well-known, but remains highly challenging to im-

plement for both patients and doctors in practice. Our results showed clearly that GlucoGuide is a novel self-management tool for people with T2D. Our clinical trial showed that GlucoGuide does help T2D patients to alleviate their diabetes conditions based on two standard clinical blood tests. Our work can also be regarded as a proof of concept in integrating data mining, mobile computing, and medical knowledge into an mobile intelligent system that can benefit people with chronic diseases, such as diabetes. We hope that our work could inspire future researchers to integrate data mining and mobile computing into similar mobile intelligent systems for other medical or health conditions.

#### REFERENCES

- [1] "Canadian diabetes association," Canadian Diabetes Association, 2014, <http://www.diabetes.ca/>.
- [2] 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.
- [3] R. Petrella, C. Lattanzio, and T. Overend, "Physical activity counseling and prescription among canadian primary care physicians," *Archives of Internal Medicine*, vol. 167, no. 16, pp. 1774–1781, 2010.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
- [8] R. Tibshirani, "Regression shrinkage and selection via lasso," *Journal of the Royal Society*, vol. 58, no. 1, pp. 267–288, 1996.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] C. X. Ling, T. Chen, Q. Yang, and J. Cheng, "Mining optimal actions for profitable crm," in *2nd IEEE International Conference on Data Mining*, 2002, pp. 767–770.
- [13] 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.
- [14] Q. Yang, J. Yin, C. X. Ling, and T. Chen, "Postprocessing decision trees to extract actionable knowledge," in *3rd IEEE International Conference on Data Mining*, 2003, pp. 685–688.