# Using Machine Learning to Predict the Future Development of Disease

Lanxin Miao*[§], Xuezhou Guo[†][§], Hasan T Abbas[§], Khalid A Qaraqe[‡], and Qammer H Abbasi[§]
*Communication Engineering, Glasgow College UESTC, Chengdu, China
[†]Electrical Engineering, Glasgow College UESTC, Chengdu, China
[‡]Department of Electrical & Computer Engineering, Texas A&M University at Qatar, Doha, Qatar 23874
[§]James Watt School of Engineering, University of Glasgow, Glasgow, Scotland G12 8QQ
Email: Qammer.Abbasi@glasgow.ac.uk

*Abstract*—**The objective of this research is to develop a long-term risk model for the development of cardiovascular disease (CVD) because of type-2 diabetes (T2D). We use the support vector machine (SVM) and the K-nearest neighbours algorithms on the dataset collected from a longitudinal study called Framingham Heart Study, to develop the prediction models. The dataset was first balanced by the Synthetic Minority Oversampling Technique algorithm. The SVM algorithm was then used to train the model, and after tuning the parameters and training for 1000 times, the average accuracy to correctly predict the prevalence of CVD due to T2D came out as $96.5\%$ and the average recall rate was $89.8\%$. Similarly, we also applied the KNN algorithm to train the dataset, and the recall rate even reaches $92.9\%$. The advantages of our model are: 1) it can predict with high accuracy both the risk of development of T2D and CVD simultaneously; 2) it can be used without the expensive and tedious oral glucose tolerance test. The model yielded high-performance results after training on the Framingham Heart Study dataset.**

*Index Terms*—**Disease prediction, diabetes, cardiovascular disease, SMOTE, SVM, Distance Correlation, Relief.**

## I. Introduction

With the improvement of people's living standards and a lax attitude towards maintaining a balanced diet as well as regular exercise, the incidence of type-2 diabetes (T2D) increased significantly all over the world from the end of the 20th century to the beginning of the 21st century. According to the San Antonio Heart Study that took place between 1987 and 1996, and in which more than 5,000 patients were enrolled, the incidence of T2D almost doubled between 7 and 8 years in both Mexican Americans and non-Hispanic people of white ethnicity [1]. Similarly, in China, the national survey in 2013 demonstrated that the prevalence of T2D was 10% [2]. Moreover, there has also been an increasing trend in the incidence of cardiovascular disease (CVD) worldwide in recent years since CVD is closely related to T2D. A piece of robust evidence that T2D was markedly associated with increased all-cause mortality and increased CVD mortality was demonstrated in [3]. Besides, a homeostasis model assessment of insulin resistance indicates a great relation in the incidence of CVD with T2D [4]. According to the annual report of the International Diabetes Federation (IDF), 12% of the global health expenditure is spent on diabetes and its complications. To reduce this huge cost, accurate disease prediction is necessary and an effective prognostic scheme must be devised which allows potential patients to have earlier treatments before progressing to more severe diseases.

Although there are many kinds of research related to the prediction of T2D, little has been done quantitatively to study the effects of T2D and CVD together. There are numerous motivation factors to carry out this research. First, disease prevention is a noble cause; better prediction models help high-risk patients to have prevention treatments in time and therefore, not only improve the quality of life but save the nation from the burden of the associated treatment costs. Secondly, the development of disease risk prediction models is rarely validated through datasets that are collected separately from the ones upon which the models were trained. In this study, we use two datasets collected in separate, independent studies. The data from the Framingham Heart Study (FHS) is used to train and test the model.

Nowadays, there is an increasing trend in the incidence of T2D [1]. Many medical organisations have been working hard to find an approach to predict T2D accurately. Several empirical indicators have been proposed to measure the risk of developing T2D, which is also called prediabetes. For example, the Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) is an index to quantify insulin resistance [4]. Matsuda index [5] is a measurement of insulin sensitivity. Traditionally, the standard model to predict T2D is through the use of the oral glucose tolerance test (OGTT), which is a blood test in which glucose is given and blood samples were taken multiple times in two hours to determine how quickly it is cleared from the blood. However, the OGTT is time-consuming and expensive. To replace this inconvenient procedure, many new models have been tested in clinical trials, including a modified insulin secretion index and a clinical model developed from the SAHS. Besides, some other traditional methods to predict T2D were proposed in [6]–[10]. Moreover, researches related to the correlation between T2D and CVD can be found in [3], [11], [4]. When it comes to utilising machine learning methods to predict diabetes or other diseases, different models have been proposed in recent years. The paper [12] used four separate machine learning algorithms to predict Diabetic Mellitus among the adult population, and the decision tree was found to provide higher accuracy. In [13], different decision tree classifiers are applied and evaluated based on their true positive rate and precision. In [14], ten features were selected from the SAHS dataset upon which SVM was to predict the future development of T2D. Although the studies produce good results in terms of prediction of future T2D, they only employ a single dataset and therefore, the results cannot be generalised for all demographics and geographical settings. Also, some serious complications such as the CVD are not discussed. In this paper, we study the effect of T2D on the development of CVD through machine learning.

## II. Methodology

Here we briefly introduce the approaches used to develop the prediction models, including data processing, feature selection, model training, and evaluation. Initially, we examined the publicly available datasets from four large institutions, including FHS, Hospital Frankfurt Diabetes Centre, and National Institute of Diabetes and Digestive and Kidney Diseases. Among them, only the FHS dataset includes cardiovascular records. Therefore, we collected the FHS dataset as a training and testing set. The FHS dataset had a total of 8,391 subjects (including 2,133 men and 6,258 women) aged 40–90 years [15].

### *A. Data Pre-processing and Feature Selection*

All the null and error values in the dataset were first deleted. To simplify the model construction and training procedure, the classification was converted into a binary scheme by assigning a positive label to the samples with both diabetes and cardiovascular disease, and negative label to other samples. Moreover, we normalised all the features to have values between 0 and 1.

The FHS dataset was highly imbalanced with the negative class being 37:3,801. As it is well known that such a high-class imbalance leads to biased results, we applied over-sampling and down-sampling techniques to balance the two classes. For over-sampling, we used the SMOTE algorithm. For down-sampling, we randomly removed samples of the majority class. After sampling the dataset, the class ratio became 200:799 in the FHS dataset.

Since the tedious OGTT is time-consuming and expensive, we selected the body mass index (BMI), age, and the fasting plasma glucose (FPG) as input features to train the SVM and KNN models.

### *B. Model Training*

To develop the risk models, we used the SVM and KNN classification algorithms separately on the FHS dataset. In terms of KNN, it is one of the simplest models to build a classifier, an object is classified by the labels of K nearest neighbours [16]. These K neighbours can do a simple majority vote and decide the category of new data. The model based on KNN is determined by three basic elements: distance measurement, K value selection, and classification decision rules.

The SVM algorithm constructs a hyperplane or set of hyperplanes in a high dimensional space for classification [17], [18]. The nearest samples to the hyper-plane are the support vectors, which influence the position and orientation of hyperplane. The distance between a hyper-plane and the support vectors is known as the functional margin. Intuitively, a good classification is achieved with the aid of a larger functional margin, which decreased generalisation error. generalisation error.

With the adoption of the SVM and KNN algorithm, we used the FHS dataset to train and test the model. For each of the 1,000 training iterations, the dataset was shuffled and divided into training and testing set at a ratio of 4:1. In the training process, we adjusted three major parameters for SVM; kernel function, C(penalty), and $\Gamma$. By using the grid-search method, the best parameter combination was selected. For the KNN model, we set parameters including n-neighbours, weights, and the metric. By evaluating the results of the trained model based on different parameters, the best parameters were found.

### *C. Evaluation*

After obtaining the developed prediction model, we evaluated it from 2 aspects. Specifically, the developed model predicts whether a person will develop CVD due to T2D according to the features. The criterion for evaluation included accuracy and recall rate (sensitivity). In most disease prediction problems, the recall rate (sensitivity) is the more reliable criterion instead of accuracy, due to the unavoidable imbalanced dataset. Recall rate refers to the true positive rate ($\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$), where TP and FN are the true positives and false negatives respectively.

## III. Results and Discussion

To train and test the model, we chose BMI, age, and FPG as the input features. Results of the KNN and SVM model are shown separately. When training the KNN model for 1,000 times in the FHS dataset, the corresponding accuracy, and recall rate are shown in Table I. The accuracy is at $96.9\%$ and the recall rate reaches $92.9\%$. The selected parameters combination for the KNN is: {`′n′:5,′weights′:′uniform′,′algorithm′: ′auto′, ′metric′:′minkowski′,′leaf-size′: 30, ′p′:2`}. The confusion matrix of the KNN model is shown in Fig. 1. Because some data were synthesised during the oversampling procedure, the confidence level in the sampled datasets decreased, however, the overall performance remained high.

TABLE I
Average accuracy and recall rate obtained from the KNN model over 1,000 iterations.

| | Result (%) |
|---|---|
| Accuracy | 96.929056 |
| Recall rate | 92.868728 |

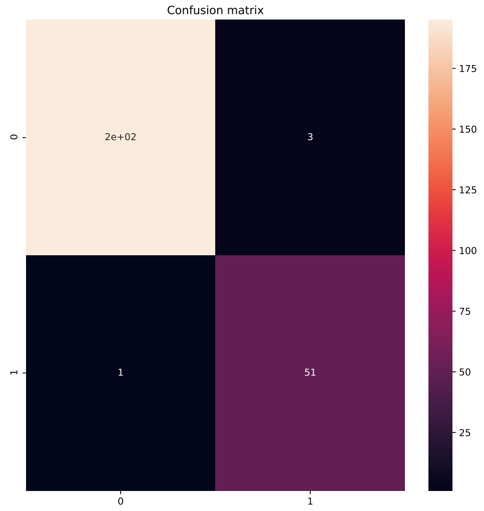

Fig. 1. The confusion matrix for the trained KNN model.

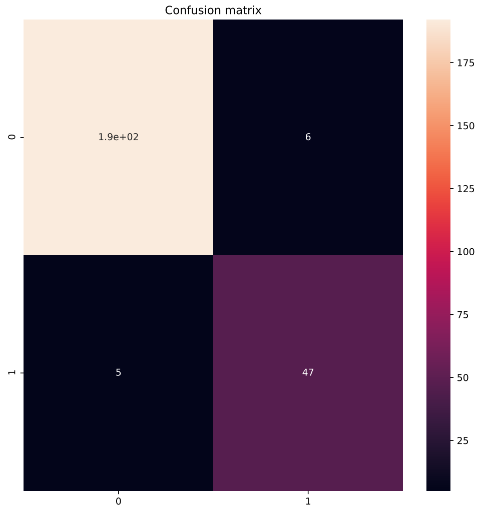

Fig. 2. The confusion matrix for the SVM model.

When training the SVM model for 1,000 times in the FHS dataset using the grid-search, the best parameters, the corresponding accuracy, and recall rate are shown in Table II. The selected best parameters set is: {`′C′:100, ′gamma′:20, ′kernel′:′rbf′`}. The model performed well with the average accuracy at 96.5%, and the sensitivity reaching 89.8%, as shown in Fig. 2. The performance in terms accuracy is less satisfactory than that of the KNN model, but the results are also persuasive in this disease prediction problem.

TABLE II
AVERAGE ACCURACY AND RECALL RATE OBTAINED FROM THE SVM MODEL OVER 1,000 ITERATIONS.

| | Result (%) |
|---|---|
| Accuracy | 96.535513 |
| Recall rate | 89.843519 |

## IV. CONCLUSIONS

In this paper, we used machine learning to develop a future cardiovascular disease (CVD) risk prediction model due to type-2 diabetes (T2D). The support vector machine (SVM) and K nearest neighbours (KNN) supervised learning algorithms were used to develop the model for which the Framingham dataset was utilised for training and testing. A remarkable aspect of this study is that it only requires anthropometric measurements and standard blood test recordings but still yield excellent results. Despite the huge imbalance of both the datasets, our model had an average accuracy calculated over 1,000 iterations equal to 96.5% and a recall rate of 89.8% in the FHS dataset based on SVM method, and more importantly, the recall rate reaches 92.9% when training KNN in the FHS dataset. In comparison with other risk prediction models, our model does not require the costly oral glucose tolerance test. Furthermore, it can simultaneously predict both the future developments of CVD and T2D.

For the future work, we recommend a multi-class classification performed on the dataset and evaluating the developed models on contemporary datasets.

## ACKNOWLEDGEMENT

This publication was made possible by NPRP grant number NPRP 10-1231-160071 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

## REFERENCES

[1] J. P. Burke, K. Williams, S. P. Gaskill, H. P. Hazuda, S. M. Haffner, and M. P. Stern, "Rapid rise in the incidence of type 2 diabetes from 1987 to 1996: results from the San Antonio Heart Study," *Archives of Internal Medicine*, vol. 159, no. 13, pp. 1450–1456, 1999.

[2] C. D. Society, "Guidelines for the prevention and treatment of type 2 diabetes in China (2017 edition)," *Chin J Diabetes*, vol. 10, no. 1, pp. 4–67, 2018.

[3] M. Wei, S. P. Gaskill, S. M. Haffner, and M. P. Stern, "Effects of diabetes and level of glycemia on all-cause and cardiovascular mortality: the San Antonio Heart Study," *Diabetes care*, vol. 21, no. 7, pp. 1167–1172, 1998.

[4] A. J. Hanley, K. Williams, M. P. Stern, and S. M. Haffner, "Homeostasis model assessment of insulin resistance in relation to the incidence of cardiovascular disease: the San Antonio Heart Study," *Diabetes care*, vol. 25, no. 7, pp. 1177–1184, 2002.

[5] M. Matsuda and R. A. DeFronzo, "Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp," *Diabetes Care*, vol. 22, pp. 1462–1470, Sep 1999.

[6] C. Lorenzo, K. Williams, and S. Haffner, "Insulin secretion based on the late oral glucose tolerance test period and incident diabetes: the San Antonio Heart Study," *Diabetic Medicine*, vol. 29, no. 8, pp. e151–e158, 2012.

[7] M. P. Stern, K. Williams, and S. M. Haffner, "Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test?," *Annals of internal medicine*, vol. 136, no. 8, pp. 575–581, 2002.

[8] M. A. Abdul-Ghani, T. Abdul-Ghani, M. P. Stern, J. Karavic, T. Tuomi, I. Bo, R. A. DeFronzo, and L. Groop, "Two-step approach for the prediction of future type 2 diabetes risk," *Diabetes Care*, vol. 34, no. 9, pp. 2108–2112, 2011.

[9] M. A. Abdul-Ghani, K. Williams, R. A. DeFronzo, and M. Stern, "What is the best predictor of future type 2 diabetes?," *Diabetes care*, vol. 30, no. 6, pp. 1544–1548, 2007.

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

[11] C. Lorenzo, K. Williams, K. J. Hunt, and S. M. Haffner, "Trend in the prevalence of the metabolic syndrome and its impact on cardiovascular disease incidence: the San Antonio Heart Study," *Diabetes care*, vol. 29, no. 3, pp. 625–630, 2006.

[12] M. F. Faruque, I. H. Sarker, *et al.*, "Performance analysis of machine learning techniques to predict diabetes mellitus," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–4, IEEE, 2019.

[13] D. Vigneswari, N. K. Kumar, V. G. Raj, A. Gugan, and S. Vikash, "Machine learning tree classifiers in predicting diabetes mellitus," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pp. 84–87, IEEE, 2019.

[14] H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani, and K. Qaraqe, "Predicting diabetes in healthy population through machine learning," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 567–570, IEEE, 2019.

[15] B. Wang, M.-C. Liu, X.-Y. Li, X.-H. Liu, Q.-X. Feng, L. Lu, Z. Zhu, Y.-S. Liu, W. Zhao, and Z.-N. Gao, "Cutoff point of HbA1c for diagnosis of diabetes mellitus in Chinese individuals," *PLOS One*, vol. 11, no. 11, p. e0166597, 2016.

[16] Tavish Srivastava, "Introduction to k-nearest neighbors: A powerful machine learning algorithm," 2018. [Online; accessed 28-August-2019].

[17] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[18] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*, pp. 11–30, Springer, 2015.