

# A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning

Ran CUI<sup>a,1</sup>, Elena DASKALAKI<sup>a</sup>, Md Zakir HOSSAIN<sup>a</sup>, Artem LENSKIY<sup>a</sup>,  
Christopher J NOLAN<sup>b</sup> and Hanna SUOMINEN<sup>a,c,d</sup>

<sup>a</sup>*School of Computing, The Australian National University (ANU), Australia*

<sup>b</sup>*ANU Medical School and John Curtin School of Medical Research, ANU, Australia*

<sup>c</sup>*Data61, Commonwealth Scientific and Industrial Research Organisation, Australia*

<sup>d</sup>*Department of Computing, University of Turku, Finland*

**Abstract.** Diabetes can be diagnosed by either Fasting Plasma Glucose or Hemoglobin A1c. The aim of our study was to explore the differences between the two criteria through the development of a machine learning based diabetes diagnostic algorithm and analysing the predictive contribution of each input biomarker. Our study concludes that fasting insulin is predictive of diabetes defined by FPG, but not by HbA1c. Besides, 28 other fasting blood biomarkers were not significant predictors of diabetes.

**Keywords.** Diabetes biomarkers, machine learning, feature importance

## 1. Introduction

Diabetes is a chronic disease which affects a large and increasing number of people worldwide. According to the *World Health Organisation* (WHO), diabetes can be diagnosed by either a *Fasting Plasma Glucose* (FPG)  $\geq 7.0$  mmol/l or a HbA1c  $\geq 6.5\%$ . Although FPG and HbA1c exhibit a proven correlation [1], the two diagnostic criteria FPG  $\geq 7.0$  mmol/l and HbA1c  $\geq 6.5\%$  are not always equivalent leading often to contradicting outcomes [2]. The aim of our study was to explore the difference of the two aforementioned diabetes diagnostic criteria through *Machine Learning* (ML) as well as investigating the contribution of 30 separate blood biomarkers<sup>2</sup> on diabetes prediction.

## 2. Approach

After obtaining ethical approval (2020/515) from the ANU Human Research Ethics Committee, we analysed longitudinal data collected in the *China Health and Nutrition Survey* (CHNS) [3], which targeted public health among eight provinces in China from 1989 to 2011. We made use of the 30 fasting blood biomarkers included in the CHNS,

---

<sup>1</sup> Corresponding Author, Ran Cui; E-mail: ran.cui@anu.edu.au.

together with gender and age information of 9,549 individuals. Among all these participants, 550 and 642 subjects satisfied  $FPG \geq 7.0$  mmol/l and  $HbA1c \geq 6.5\%$ , respectively. We formulated the diagnosis of diabetes as a ML classification task tackled with the XGBoost algorithm, which took the biomarkers, gender, and age as inputs. We applied stratified sampling to balance the number of positive (diabetes) and negative (non-diabetes) samples in the experiments. To better investigate the relationship between FPG and HbA1c, we designed the following two experiments: 1) excluding both FPG and HbA1c in the input features set, while using  $FPG \geq 7.0$  mmol/l or  $HbA1c \geq 6.5\%$  as the definition of diabetes in separate analyses (experiment 1), and including FPG or HbA1c in the input features set when the other was used as the ground truth label (experiment 2). We then employed permutation feature importance [4] to study the contribution of each feature (biomarker) towards the final prediction of diabetes diagnosis.

### 3. Results

Performance, measured by the F1 score, dropped when both FPG and HbA1c were removed from the input biomarkers set (Table 1). This agrees with the aforementioned proven correlation of FPG and HbA1c and their utility for diabetes diagnosis. Although fasting insulin (INS) was also an important predictive biomarker (Table 2) when diabetes was diagnosed by  $FPG \geq 7.0$  mmol/l, it was not predictive of diabetes diagnosed by HbA1c. Moreover, age possessed high importance in our models, which showed its strong position as a diabetes risk factor. None of the remaining biomarkers were strongly correlated with the presence of diabetes, which suggested that they were not informative and can be safely discarded from our model without affecting algorithmic performance.

**Table 1.** F1 scores in our experiments.

Diabetes definition	Experiment 1	Experiment 2
$FPG \geq 7.0$ mmol/l	0.822	0.893
$HbA1c \geq 6.5\%$	0.791	0.867

**Table 2.** Sorted average permutation features importance [%] in the two experiments. Features (biomarkers) that were significant have been bolded.

Diabetes Definition	Experiment 1		Experiment 2		Diabetes Definition	Experiment 1		Experiment 2	
$FPG \geq 7.0$ mmol/l	<b>INS</b>	<b>2.85</b>	<b>HbA1c</b>	<b>6.96</b>	$HbA1C \geq 6.5\%$	<b>age</b>	<b>1.40</b>	<b>FPG</b>	<b>7.04</b>
	<b>age</b>	<b>1.42</b>	<b>INS</b>	<b>3.32</b>		MG	0.59	ALT	0.31
	ALB	0.57	age	0.40		APO_B	0.48	TP	0.29
	UA	0.52	Y48_3	0.28		HS_CRP	0.41	APO_A	0.29
	Y48_4	0.45	APO_A	0.17		AP0_A	0.37	Age	0.28
	MG	0.39	Y48_4	0.13		FET	0.34	RBC	0.28
	...	...	...	...		...	...	...	...
	TRF_R	-0.05	Y48_5	-0.22		LP_A	-0.27	UA	-0.03
	LDL_C	-0.07	Y48_2	-0.22		TRF	-0.35	PLT	-0.09

#### 4. Conclusions

We conclude our study with two clinical suggestions: (1) Our finding of fasting insulin being an important biomarker only when diabetes is defined by FPG, but not by HbA1c, suggests that HbA1c should be considered in conjunction with insulin levels in fasting blood samples to ensure an accurate diagnosis, while this does not hold if the diagnosis is based on FPG. (2) Most biomarkers other than FPG, HbA1c, and fasting insulin included in our experiments were not significantly more important than age, hence a clinician could refer less to them when predicting diagnosis of diabetes.

#### Acknowledgements

This research has been delivered in partnership with and funded by *Our Health in Our Hands* (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. We acknowledge the funding from the ANU School of Computing for the first author's PhD studies.

#### References

- [1] Khan H, Sobki S, Khan S. Association between glycaemic control and serum lipids profile in type 2 diabetic patients: HbA1c predicts dyslipidaemia. *Clinical and Experimental Medicine* 7 (2007), 24–29.
- [2] Sacks DB. A1C versus glucose testing: a comparison. *Diabetes Care* 34 (2011), 518–523.
- [3] Popkin BM, Du S, Zhai F, Zhang B. Cohort Profile: The China Health and Nutrition Survey—monitoring and understanding socio-economic and health change in China, 1989–2011. *International Journal of Epidemiology* 39 (2009), 1435–1440.
- [4] Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *arXiv, arXiv preprint arXiv:1801.01489* (2018).