Table SI. Sequences of the primers used in the present study.

Species	Identification of the reference genome	Direction	Primer sequences (5'-3')
Veillonellaceae	MZ093067.1	Forward	CCCGGGCCTTGTACACACCG
		Reverse	CCCACCGGCTTTGGGCACTT
Clostridium leptum	MT903091	Forward	GCACAAGCAGTGGAGT
		Reverse	CTTCCTCCGTTTTGTCAA
Roseburia inulinivorans	AB661436.1	Forward	TCTGACCGGACAGTAATGTG
		Reverse	CGCTGGCTACTGGGGATAAG
Bacteroides	OK326644.1	Forward	CTGAACCAGCCAAGTAGCG
		Reverse	CCGCAAACTTTCACAACTGACTTA
Prevotella	OL906056.1	Forward	CCAGCCAAGTAGCGTGCA
		Reverse	TGGACCTTCCGTATTACCGC
Bifidobacterium	OL677371.1	Forward	TCGCGTCCGGTGTGAAAG
		Reverse	CCACATCCAGCATCCAC
Lactobacillus	OK655764.1	Forward	AGCAGTAGGGAATCTTCCA
		Reverse	CACCGCTACACATGGAG
Faecalibacterium prausnitzii	OK510341.1	Forward	CCCTTCAGTGCCGCAGT
		Reverse	GTCGCAGGATGTCAAGAC
Enterococcus	OL872201.1	Forward	CCCTTATTGTTAGTTGCCATCATT
		Reverse	ACTCGTTGTACTTCCCATTGT
Eubacterium rectale	MT903123.1	Forward	GGAATATTGCACAATGGGC
		Reverse	AGCCGGTGCTTCTTAGTCAG
Internal control (16S rDNA)		Forward	CGTCAGCTCGTGYCGTGAG
		Reverse	CGTCRTCCCCRCCTTCC

Table SII. Selection of parameters in the predictive models.

Models	Tuning parameter
SVM	C (Regularization parameter) Gamma (Kernel coefficient) kernel (kernel type) class_weight
MLP	α (Regularization parameter) hidden_layer_sizes (Network architecture) max_iter (Maximum number of iterations)
XGboost	colsample_bytree (Subsample ratio of columns) gamma (Minimum loss reduction) max_depth (Maximum depth of a tree) min_child_weight (Minimum sum of instance weight needed in a child) n_estimators (The number of rounds for boosting) reg_alpha (L1 regularization term) reg_lambda (L2 regularization term) subsample (Subsample ratio of the training instances)

SVM, support vector machine; MLP, multilayer perceptron.

Table SIII. Clinical characteristics of patients with T2DM and control subjects.

Clinical characteristics	Control (n=89)	Patients with T2DM (n=118)	P-value
Age, years (median and range)	57 (27-85)	59 (35-84)	0.502
Sex			0.683
Male	48	67	
Female	41	51	
Body mass index (kg/m^2) (median and range)	25.01 (18.37-34.05)	25.56 (18.34-32.66)	0.23
Hypertension			0.055
Yes	60	64	
No	29	54	
Smoking			0.146
Yes	35	35	
No	54	83	
Alcohol consumption			0.22
Yes	32	33	
No	57	85	

T2DM, type 2 diabetes mellitus.

Table SIV. Summary of the efficacy of the SVM, MLP and XGboost models in the training set.

Metrics	SVM (95% CI)	XGboost (95% CI)	MLP (95% CI)
Accuracy	0.63 (0.52-0.75)	0.62 (0.55-0.75)	0.64 (0.55-0.78)
Precision	0.71 (0.58-0.83)	0.67 (0.6-0.75)	0.67 (0.6-0.77)
Recall	0.63 (0.53-0.73)	0.67 (0.58-0.74)	0.72 (0.61-0.88)
Specificity	0.64 (0.44-0.85)	0.56 (0.37-0.71)	0.53 (0.43-0.64)
Npv	0.56 (0.44-0.68)	0.56 (0.46-0.64)	0.6 (0.47-0.8)

SVM, support vector machine; MLP, multilayer perceptron; 95% CI, 95% confidence interval; Npv, negative prognostic value.

Table SV. Summary of the efficacy of the SVM, MLP and XGboost models in the test set.

Metrics	SVM	XGboost	MLP
Accuracy	0.69	0.67	0.67
Precision	0.79	0.75	0.78
Recall	0.62	0.62	0.58
Specificity	0.78	0.72	0.78
Npv	0.61	0.59	0.58

SVM, support vector machine; MLP, multilayer perceptron; 95% CI, 95% confidence interval; Npv, negative prognostic value.