Contents lists available at ScienceDirect

Diabetes Epidemiology and Management

journal homepage: www.elsevier.com

Original article

Predictive modeling for the development of diabetes mellitus using key factors in various machine learning approaches

Marenao Tanaka[a,b,1], Yukinori Akiyama[c,1], Kazuma Mori[d], Itaru Hosaka[e], Kenichi Kato[e], Keisuke Endo[a], Toshifumi Ogawa[a,f], Tatsuya Sato[a,f], Toru Suzuki[a,g], Toshiyuki Yano[a], Hirofumi Ohnishi[h], Nagisa Hanawa[i], Masato Furuhashi[a,*]

[a] *Department of Cardiovascular, Renal and Metabolic Medicine, Sapporo Medical University School of Medicine, S-1, W-16, Chuo-ku, Sapporo 060-8543, Japan*
[b] *Tanaka Medical Clinic, Yoichi, Japan*
[c] *Department of Neurosurgery, Sapporo Medical University, Sapporo, Japan*
[d] *Department of Immunology and Microbiology, National Defense Medical College, Tokorozawa, Japan*
[e] *Department of Cardiovascular Surgery, Sapporo Medical University School of Medicine, Sapporo, Japan*
[f] *Department of Cellular Physiology and Signal Transduction, Sapporo Medical University School of Medicine, Sapporo, Japan*
[g] *Natori Toru Internal Medicine and Diabetes Clinic, Natori, Japan*
[h] *Department of Public Health, Sapporo Medical University School of Medicine, Sapporo, Japan*
[i] *Department of Health Checkup and Promotion, Keijinkai Maruyama Clinic, Sapporo, Japan*

ARTICLE INFO

Article History:
Received 27 November 2023
Revised 30 November 2023
Accepted 11 December 2023
Available online 14 December 2023

Keywords:
Artificial intelligence
Machine learning
Diabetes mellitus
Fatty liver index

ABSTRACT

Aims: Machine learning (ML) approaches are beneficial when automatic identification of relevant features among numerous candidates is desired. We investigated the predictive ability of several ML models for new onset of diabetes mellitus.
Methods: In 10,248 subjects who received annual health examinations, 58 candidates including fatty liver index (FLI), which is calculated by using waist circumference, body mass index and levels of triglycerides and γ-glutamyl transferase, were used.
Results: During a 10-year follow-up period (mean period: 6.9 years), 322 subjects (6.5 %) in the training group (70 %, n=7,173) and 127 subjects (6.2 %) in the test group (30 %, n=3,075) had new onset of diabetes mellitus. Hemoglobin A1c, fasting glucose and FLI were identified as the top 3 predictors by random forest feature selection with 10-fold cross-validation. When hemoglobin A1c and FLI were used as the selected features, C-statistics analogous in receiver operating characteristic curve analysis in ML models including logistic regression, naïve Bayes, extreme gradient boosting and artificial neural network were 0.874, 0.869, 0.856 and 0.869, respectively. There was no significant difference in the discriminatory capacity among the ML models.
Conclusions: ML models incorporating hemoglobin A1c and FLI provide an accurate and straightforward approach for predicting the development of diabetes mellitus.

Introduction

Diabetes mellitus is a critical issue in modern public health since it compromises duration of life and healthy longevity [1,2]. The prevalence of diabetes mellitus is continuously increasing worldwide, and major causes of death including cardiovascular disease, stroke, renal disfunction and cancer are triggered by diabetes mellitus [3,4]. Hence, the establishment of methods for precise prediction of the development of diabetes mellitus is needed and it is important for public health and clinical environment [5].

Obesity, genetic factor, stress, comorbidities of hypertension and dyslipidemia, and lifestyles such as smoking habits, excess consumption of alcohol and the lack of exercise are associated with the development of diabetes mellitus [2]. It has been reported that liver dysfunction is also a risk factor of diabetes mellitus [6,7]. We previously reported that hazard risk for the development of diabetes mellitus increases in subjects with high levels of alanine aminotransferase (ALT) and γ-glutamyl transferase (GGT) [8]. Steatotic liver diseases including nonalcoholic fatty liver disease (NAFLD), metabolic dysfunction-related fatty liver disease (MAFLD) and metabolic dysfunction-associated steatotic liver disease (MASLD) are associated with lifestyle-related diseases such as metabolic syndrome and diabetes mellitus [9–12]. It has been reported that fatty liver index (FLI), which is calculated by using body mass index (BMI), waist

[*] Corresponding author.
E-mail address: furuhasi@sapmed.ac.jp (M. Furuhashi).
[1] The first two authors equally contributed to this work.

https://doi.org/10.1016/j.deman.2023.100191

circumference (WC) and levels of GGT and triglycerides (TG), is a noninvasive and simple predictor for hepatosteatosis [13], NAFLD [14] and MAFLD [10] and that a high level of FLI can predict new onset of diabetes mellitus [15], chronic kidney disease[16], hypertension [17] heart failure [18], and ischemic heart disease [19].

Recently, machine learning (ML) programs have been used in social community widespread and have been widely applied in medical fields [20]. Although classically statistical methods often require manual feature selection by researchers, ML methods are beneficial when dealing with feature selection and extraction tasks are difficult or when automatic identification of relevant features is desired among numerous candidates [21,22]. In the present study, we investigated the importance of various features at baseline for the development of diabetes mellitus during a 10-year follow-up period in several ML models.

Methods

The present study was a retrospective and single-center cohort study conducted in Japan. The study conformed to the principles outlined in the Declaration of Helsinki and was conducted with the approval of the Ethics Committee of Sapporo Medical University (Number: 30−2−32). Written informed consent was obtained from all of the subjects.

Study subjects

All of the subjects who received annual health examinations at Keijinkai Maruyama Clinic, Sapporo, Japan in 2006 were enrolled in this registry (n = 28,990). A flow chart of the study participants in a discovery study is shown in Supplementary Fig. S1. Prespecified exclusion criteria were the absence of data for fasting glucose and hemoglobin A1c and diagnosis of diabetes mellitus at baseline. After exclusion, a total of 11,653 subjects who received health examinations at least once in the period from 2007 to 2016 were included in the discovery study.

After identifying important factors for the development of diabetes mellitus in the discovery study, various ML models using the selected features were constructed in a modeling study. In the modeling study, prespecified exclusion criteria were the absence of data for fasting glucose, hemoglobin A1c and components of FLI calculation including BMI, WC, GGT and TG and diagnosis of diabetes mellitus at baseline (Fig. 1). After exclusion, a total of 10,248 subjects who received health examinations at least once in the period from 2007 to 2016 were included. In both the discovery and modeling studies, the enrolled subjects were randomly divided into a training group (70 %) and a test group (30 %) using the train_test_split function in scikit-learn (1.0.2) in Python libraries.

A self-administered questionnaire survey was performed to obtain information on current smoking habit, alcohol drinking habit (≥ 3 times/week), treatment of hypertension, diabetes mellitus, dyslipidemia, hyperuricemia and ischemic heart disease, and family history of hypertension and diabetes mellitus.

Clinical endpoint

The clinical endpoint was the development of diabetes mellitus during a 10-year follow-up period. Diabetes mellitus was diagnosed in accordance with the guidelines of the American Diabetes Association [1]: fasting glucose ≥ 126 mg/dL, hemoglobin A1c ≥ 6.5 % or self-reported use of anti-diabetic drugs.

Measurements

Medical examinations including samplings of urine and blood were performed after an overnight fast. Blood pressure was measured on the arm using a sphygmomanometer (#601, Kenzmedico, Saitama, Japan). BMI was calculated as body weight in kilograms divided by squared height in meters. Levels of total cholesterol, high-density lipoprotein (HDL) cholesterol and TG were measured by enzymatic assays. Non-HDL cholesterol level was calculated by subtracting HDL cholesterol level from total cholesterol level. Low-density lipoprotein (LDL) cholesterol, large buoyant LDL cholesterol and small dense LDL cholesterol were calculated by using the Sampson's equation [23,24]: LDL cholesterol [23] = total cholesterol/0.948 − HDL cholesterol/0.971 − (TG/8.56 + [TG × non-HDL cholesterol]/2140 − TG^2/16,100) − 9.44; large buoyant LDL cholesterol [24] = 1.43 × LDL cholesterol − (0.14 × (In [TG] × LDL cholesterol) − 8.99; small dense LDL cholesterol [24] = LDL cholesterol − large buoyant LDL cholesterol. Estimated glomerular filtration rate (eGFR) was calculated by the following equation for Japanese people [25]: eGFR (mL/min/1.73m^2) = 194 × serum creatinine $^{(-1.094)}$ × age $^{(-0.287)}$ × 0.739 (if female). FLI was calculated by using the equation reported by Bedogni et al. [13] FLI = $[e^{(0.953 \times \ln (TG) + 0.139 \times BMI + 0.718 \times \ln (GGT) + 0.053 \times WC-15.745)}]/[1 + e^{(0.953 \times \ln (TG) + 0.139 \times BMI + 0.718 \times \ln (GGT) + 0.053 \times WC -15.745)}] \times 100$. FIB-4 index, a marker of hepatic fibrosis, was calculated by the following formula [26]: age × aspartate aminotransferase (AST)/(platelet count [10^9/L] × $ALT^{1/2}$).

Statistical analysis

Numeric variables are expressed as means ± standard deviation (SD) for parameters with normal distributions and as medians (interquartile ranges) for parameters with skewed distributions. Intergroup differences in percentages of demographic parameters were examined by the chi-square test. The distribution of each parameter was tested for its normality using the Shapiro-Wilk W test. Comparisons between two groups for parametric and nonparametric factors were performed by using Student's *t*-test and the Mann-Whitney U test, respectively.

In ML models, missing values in parameters with normal distributions and those with skewed distributions were complemented by mean and median values, respectively. To select the best predictors for the development of diabetes mellitus, the random forest feature selection by 10−fold cross−validation was applied. The feature importance was evaluated using two metrics: mean decrease accuracy (MDA) and mean decrease Gini impurity (MDG). MDA measures the decline in prediction accuracy when a particular feature is excluded, while MDG quantifies the contribution of each feature to the uniformity of nodes and leaves in the random forest model [21,22]. Prediction for the development of diabetes mellitus was investigated by ML models using logistic regression and naïve Bayes with scikit-learn (1.0.2), extreme gradient boosting with xgboost (1.7.5), and artificial neural network with tensor flow (2.12.0) in Python libraries.

To compare the discrimination for development of diabetes mellitus between ML models, C-statistics analogous to the area under the curve (AUC) in receiver operating characteristic curve (ROC) analysis was investigated by using the method of DeLong et al [27]. A p value of less than 0.05 was considered statistically significant was defined as a meaningful difference. All data were analyzed by using EZR [28] and R version 3.6.1 and Python version 3.10.7.

Results

The discovery study

Basal characteristics of the included subjects in the discovery study are shown in Supplementary Table S1. The numbers of missing values in variables were from 1 to 2977. There were no significant differences in all variables between the training and test groups. The median numbers of health examination during the follow-up period

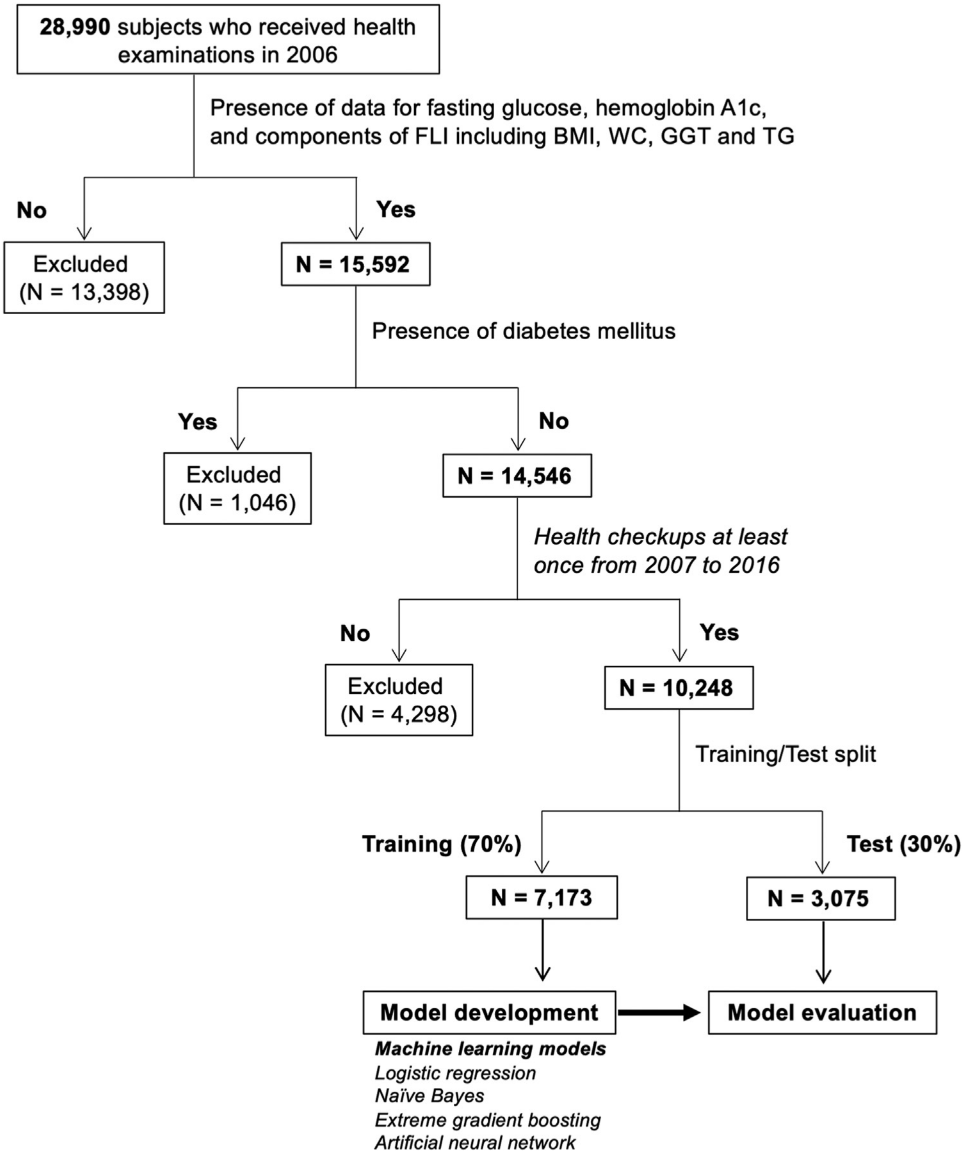
28,990 subjects who received health examinations in 2006
Presence of data for fasting glucose, hemoglobin A1c, and components of FLI including BMI, WC, GGT and TG
No
Yes
Excluded (N = 13,398)
N = 15,592
Presence of diabetes mellitus
Yes
No
Excluded (N = 1,046)
N = 14,546
Health checkups at least once from 2007 to 2016
No
Yes
Excluded (N = 4,298)
N = 10,248
Training/Test split
Training (70%)
Test (30%)
N = 7,173
N = 3,075
Model development
Model evaluation
Machine learning models
Logistic regression
Naïve Bayes
Extreme gradient boosting
Artificial neural network

Fig. 1. Flow chart of the selected study participants in the modeling study.

Among 28,990 subjects enrolled in 2006, a total of 10,248 subjects were finally included for analyses in the modeling study. The enrolled subjects were randomly divided into a training group (70 %, n = 7173) and a test group (30 %, n = 3075).

BMI, body mass index; FLI, fatty liver index; GGT, γ-glutamyl transferase; TG, triglycerides; WC, waist circumference.

were 6 times, and the percentages of subjects who received a health examination every year from 2006 to 2016 in the training and test groups were 25.1 % and 24.6 %, respectively (Supplementary Table S2). The cumulative incidences of diabetes mellitus in the training and test groups were 6.4 % and 6.9 %, respectively (Supplementary Table S3). In random forest feature selection with 10-fold cross-validation, top 3 predictors were fasting glucose, hemoglobin A1c and FLI among 58 candidates (Supplementary Fig. S2). In addition, components of FLI calculation including WC, BMI, GGT and TG were identified among the top 10 predictors.

The modeling study

Since FLI was identified as an important factor for the development of diabetes mellitus in the discovery study (Supplementary Fig. S2), subjects with the absence of data for components of FLI calculation including BMI, WC, GGT and TG were excluded in the modeling study. After exclusion, a total of 10,248 subjects were enrolled (Fig. 1). Basal characteristics of the enrolled and excluded subjects in the modeling study are shown in Supplementary Table S4. The excluded subjects were significantly younger than the enrolled subjects and had higher levels of fasting glucose and hemoglobin A1c and larger BMI than did the enrolled subjects. There were no significant differences in levels of FLI, WC, GGT and TG between the excluded and enrolled subjects.

Characteristics of the enrolled subjects in the modeling study are shown in Table 1. There were no significant differences in all variables between the training and test groups. Median values of FLI in the training and test groups were both 19 in the modeling study (Table 1), which was same as the results in the discovery study (Supplementary Table S1). The numbers of subjects who received annual health examinations during a 10-year period are shown in Supplementary Table S2. The mean follow-up period was 6.9 years (range: 1 to 10 years), and the median numbers of health examination during the follow-up period were 6 times in the modeling study. The percentages of subjects who received a health examination every year from

Table 1
Characteristics of the recruited subjects at baseline in the modeling study (n = 10,248).

	Defect n	Total n = 10,248	Training n = 7173	Test n = 3075	P
Age, years	0	47 ± 10	47 ± 10	47 ± 10	0.911
Sex, Men	0	6362 (62.1)	4472 (62.3)	1890 (61.5)	0.399
BMI	0	22.8 ± 3.2	22.8 ± 3.2	22.8 ± 3.2	0.679
Body surface area, m^2	0	1.69 ± 0.18	1.69 ± 0.19	1.69 ± 0.18	0.742
WC, cm	0	82.4 ± 9.0	82.4 ± 9.0	82.4 ± 8.9	0.965
Systolic blood pressure, mmHg	0	112 ± 13	112 ± 13	112 ± 13	0.165
Diastolic blood pressure, mmHg	0	72 ± 9	71 ± 9	71 ± 9	0.513
Heart rate, per minute	62	63 ± 9	63 ± 9	63 ± 9	0.966
Current smoking habit	0	3637 (35.5)	2519 (35.1)	1118 (36.4)	0.233
Brinkman index	0	0 [0−320]	0 [0−300]	0 [0−320]	0.854
Alcohol drinking habit	0	4468 (43.6)	3157 (44.0)	1311 (42.6)	0.200
Treatment					
Hypertension	0	27 (0.3)	19 (0.3)	8 (0.3)	1.000
Dyslipidemia	0	269 (2.6)	197 (2.7)	72 (2.3)	0.252
Hyperuricemia	0	85 (0.8)	67 (0.9)	18 (0.6)	0.076
Ischemic heart disease	0	79 (0.8)	60 (0.8)	19 (0.6)	0.269
Family history					
Hypertension	0	2422 (23.6)	1668 (23.3)	754 (24.5)	0.171
Diabetes mellitus	0	1786 (17.4)	1268 (17.7)	518 (16.8)	0.307
Biochemical data					
Hemoglobin, g/dL	0	14.3 ± 1.6	14.3 ± 1.5	14.2 ± 1.6	0.843
Hematocrit, %	0	44 ± 4	44 ± 4	44 ± 4	0.725
Platelet, x $10^4/\mu L$	0	23.8 ± 5.2	23.8 ± 5.2	23.8 ± 5.2	0.921
White blood cell, /μL	0	5858 ± 1718	5911 ± 1711	5910 ± 1732	0.420
Total protein, g/dL	1174	7.2 ± 0.4	7.2 ± 0.4	7.2 ± 0.3	0.448
Albumin, g/dL	1677	4.4 ± 0.2	4.4 ± 0.2	4.4 ± 0.2	0.457
Blood urea nitrogen, mg/dL	1003	14 ± 4	14 ± 3	14 ± 3	0.936
Creatinine, mg/dL	256	0.73 ± 0.22	0.73 ± 0.20	0.73 ± 0.26	0.942
eGFR, mL/min/$1.73m^2$	256	85 ± 14	85 ± 14	86 ± 14	0.697
Uric acid, mg/dL	11	5.4 ± 1.4	5.4 ± 1.4	5.4 ± 1.4	0.940
AST, U/L	0	21 [18−25]	21 [18−25]	21 [17−25]	0.190
ALT, U/L	0	20 [15−30]	20 [15−30]	20 [15−30]	0.490
GGT, U/L	0	28 [18−51]	28 [18−51]	28 [18−50]	0.283
Alkaline phosphatase, U/L	796	204 ± 60	204 ± 58	203 ± 56	0.210
Total bilirubin, mg/dL	1248	0.9 ± 0.3	0.9 ± 0.3	0.9 ± 0.3	0.421
Amylase, U/L	1949	101 ± 33	102 ± 31	101 ± 30	0.284
LDH, U/L	1308	169 ± 28	169 ± 26	169 ± 26	0.932
Cholinesterase, U/L	1949	5321 ± 1067	5348 ± 793	5345 ± 806	0.810
Fasting glucose, mg/dL	0	89 ± 9	90 ± 8	89 ± 9	0.158
Hemoglobin A1c, %	0	5.2 ± 0.4	5.2 ± 0.4	5.2 ± 0.4	0.554
Total cholesterol, mg/dL	0	204 ± 34	204 ± 34	204 ± 34	0.465
LDL cholesterol, mg/dL	3	123 ± 32	123 ± 31	123 ± 32	0.912
Small dense LDL cholesterol, mg/dL	3	34 ± 13	34 ± 13	35 ± 13	0.210
HDL cholesterol, mg/dL	2	61 ± 16	62 ± 16	61 ± 15	0.062
Non-HDL cholesterol, mg/dL	2	143 ± 35	143 ± 36	143 ± 36	0.799
Triglycerides, mg/dL	0	88 [61−130]	87 [61−130]	90 [62−130]	0.102
Serum Sodium, mEq/L	2206	142 ± 2	142 ± 1	142 ± 1	0.930
Serum Chloride, mEq/L	2206	104 ± 2	104 ± 2	104 ± 2	0.254
Serum Calcium, mEq/L	2518	9.0 ± 0.4	9.1 ± 0.3	9.1 ± 0.3	0.949
Serum Potassium, mEq/L	2206	4.0 ± 0.3	4.0 ± 0.3	4.0 ± 0.3	0.522
Serum Iron, μg/dL	2433	118 ± 39	118 ± 39	118 ± 39	0.580
Thyroid stimulating hormone, μU/mL	2637	1.6 [1.2−2.3]	1.6 [1.3−2.4]	1.6 [1.2−2.0]	0.686
C-reactive protein, mg/dL	1834	0.05 [0.05−0.07]	0.05 [0.05−0.07]	0.05 [0.05−0.07]	0.472
Urinalysis					
Proteinuria	0	344 (3.4)	244 (3.4)	100 (3.3)	0.720
Hematuria	0	1094 (10.7)	742 (10.3)	352 (11.4)	0.101
Glycosuria	0	13 (0.1)	7 (0.1)	6 (0.2)	0.229
Ketonuria	0	331 (3.2)	237 (3.3)	94 (3.1)	0.541
Urine pH	0	5.90 ± 0.75	5.90 ± 0.75	5.88 ± 0.75	0.378
Urine specific gravity	0	1.02 ± 0.01	1.02 ± 0.01	1.02 ± 0.01	0518
Biological index					
FLI	0	19 [7−43]	19 [7−43]	19 [7−43]	0.786
FIB-4 index	0	0.9 [0.7−1.2]	0.9 [0.7−1.2]	0.9 [0.7−1.2]	0.747

Variables are expressed as number (%), means ± SD or medians [interquartile ranges].
ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; eGFR, estimated glomerular filtration rate; FLI, fatty liver index; GGT, γ-glutamyl transpeptidase; HDL, high-density lipoprotein; LDH, lactate dehydrogenase; LDL, low-density lipoprotein; WC, waist circumference.

2006 to 2016 in the training and test groups were 25.1 % and 23.3 %, respectively. The cumulative incidences of diabetes mellitus in the training and test groups of the modeling study were 6.5 % and 6.2 %, respectively (Supplementary Table S3).

Feature selection in the modeling study

In the modeling study, top 3 predictors including fasting glucose, hemoglobin A1c and FLI were identified by random forest feature

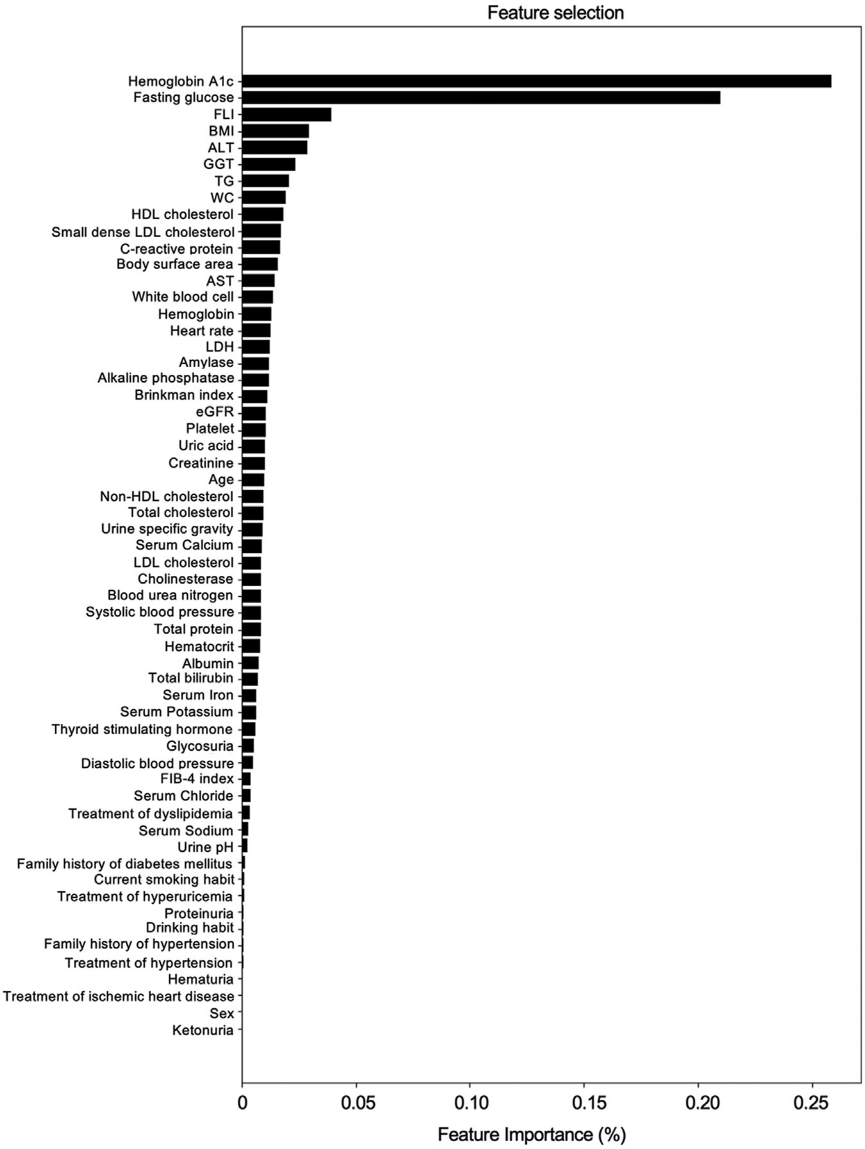

Fig. 2. Feature selection in the modeling study.

Features with high importance determined by the random forest feature selection using 10−fold cross−validation were arranged in the descending order among 58 candidates in the modeling study.

ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; eGFR, estimated glomerular filtration rate; FLI, fatty liver index; GGT, γ-glutamyl transpeptidase; HDL, high-density lipoprotein; LDH, lactate dehydrogenase; LDL, low-density lipoprotein; TG, triglycerides; WC, waist circumference.

selection with 10-fold cross-validation among 58 candidates (Fig. 2). The feature importance of components of FLI calculation including BMI, WC, GGT and TG and a liver dysfunction-related marker, ALT, were followed by the top 3 features. Since three is an apparent multicollinearity between hemoglobin A1c and fasting glucose, ML models using FLI (or components of FLI) with either hemoglobin A1c or fasting glucose were separately analyzed to predict the development of diabetes mellitus.

Discriminatory capacity of ML models for the development of diabetes mellitus using FLI

In discriminatory capacities for the development of diabetes mellitus evaluated by ROC analyses using the selected features including hemoglobin A1c and FLI, the AUCs of ML models including logistic regression, naïve Bayes, extreme gradient boosting and artificial neural network were 0.874, 0.869, 0.856 and 0.869, respectively (Table 2, Model 1). In the logistic regression model, prediction equation for the development of diabetes mellitus by was ‘p = 1 / 1 + exp (−[3.871 × hemoglobin A1c + 0.020 × FLI − 24.898])’ in Model 1. When the logistic regression model was used as the reference, discriminatory capacities for predicting development of diabetes mellitus in AUCs were not significantly different among the other ML models.

On the other hand, the AUCs in the models of logistic regression, naïve Bayes, extreme gradient boosting and artificial neural network using the features including fasting glucose and FLI were 0.839, 0.838, 0.827 and 0.837, respectively (Table 2, Model 2). There were no significant differences in AUCs of logistic regression model with the models of naïve Bayes, extreme gradient boosting and artificial neural network. In the logistic regression model, prediction equation for the development of diabetes mellitus by was ‘p = 1 / 1 + exp (−[0.126 × fasting glucose + 0.018 × FLI − 15.672])’ in Model 2.

The values of accuracy in models of logistic regression and artificial neural network, the value of sensitivity in naïve Bayes model,

Table 2
Discrimination of machine learning models for the development of diabetes mellitus during a 10-year follow-up period.

	FLI-based models				FLI components-based models			
Machine learning models	Model 1		Model 2		Model 3		Model 4	
	AUC (95 % CI)	*P*	AUC (95 % CI)	*P*	AUC (95 % CI)	*P*	AUC (95 % CI)	*P*
Logistic regression	0.874 (0.841, 0.908)	–	0.839 (0.800, 0.877)	–	0.873 (0.840, 0.907)	–	0.842 (0.804, 0.880)	–
Naïve Bayes	0.869 (0.834, 0.904)	0.128	0.838 (0.799, 0.878)	0.885	0.836 (0.799, 0.873)	< 0.001	0.816 (0.775, 0.857)	0.015
Extreme gradient boosting	0.856 (0.818, 0.894)	0.086	0.827 (0.788, 0.866)	0.262	0.849 (0.813, 0.885)	0.015	0.821 (0.779, 0.863)	0.062
Artificial neural network	0.869 (0.833, 0.904)	0.292	0.837 (0.798, 0.876)	0.612	0.871 (0.837, 0.904)	0.694	0.829 (0.787, 0.870)	0.077

Model 1: hemoglobin A1c and fatty liver index (FLI).
Model 2: fasting glucose and FLI.
Model 3: hemoglobin A1c, body mass index (BMI), waist circumference (WC), γ-glutamyl transpeptidase (GGT) and triglycerides (TG).
Model 4: fasting glucose, BMI, WC, GGT and TG.
P for logistic regression model. AUC, area under the curve; CI, confidence interval.

Table 3
Evaluation of machine learning models for the development of diabetes mellitus during a 10-year follow-up period.

	FLI-based models						FLI components-based models					
Machine learning models	Model 1			Model 2			Model 3			Model 4		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Logistic regression	0.963	0.588	0.831	0.958	0.568	0.755	0.961	0.580	0.791	0.960	0.576	0.776
Naïve Bayes	0.961	0.647	0.760	0.958	0.663	0.617	0.937	0.672	0.636	0.936	0.686	0.638
Extreme gradient boosting	0.960	0.602	0.765	0.956	0.586	0.733	0.960	0.590	0.756	0.960	0.587	0.756
Artificial neural network	0.963	0.607	0.802	0.958	0.578	0.721	0.961	0.591	0.791	0.960	0.579	0.759

Model 1: hemoglobin A1c and fatty liver index (FLI). Model 2: fasting glucose and FLI. Model 3: hemoglobin A1c, body mass index (BMI), waist circumference (WC), γ-glutamyl transpeptidase (GGT) and triglycerides (TG). Model 4: fasting glucose, BMI, WC, GGT and TG.

and the values of specificity in the logistic regression model were highest among ML models in Model 1 using hemoglobin A1c and FLI and Model 2 using fasting glucose and FLI (Table 3).

Discriminatory capacity of ML models for the development of diabetes mellitus using components of FLI

In Model 3 using hemoglobin A1c and components of FLI calculation including BMI, WC, GGT and TG instead of FLI itself, models of naïve Bayes and extreme gradient boosting had a significant deterioration of AUCs for predicting the development of diabetes mellitus compared with AUC in logistic regression model (Table 2). There was no significant difference in AUCs between models of logistic regression and artificial neural network. In Model 4 using fasting glucose, BMI, WC, GGT and TG, AUC in the naïve Bayes model was significantly lower than that in logistic regression model. There was no significant difference in AUCs between logistic regression model and models of extreme gradient boosting and artificial neural network.

The values of accuracy in models of logistic regression and artificial neural network, the value of sensitivity in naïve Bayes model, and the value of specificity in models of logistic regression and artificial neural network were highest among ML models in Model 3 (Table 3). On the other hand, the values of accuracy in models of logistic regression, extreme gradient boosting and artificial neural network, the value of sensitivity in naïve Bayes model, and the value of specificity in logistic regression model were highest among ML models in Model 4 (Table 3).

Discussion

The present study unveiled a robust predictive capability for the development of diabetes mellitus by integrating FLI with hemoglobin A1c or fasting glucose and also underscored the significance of components of FLI calculation including BMI, WC, GGT, and TG and a liver-related marker, ALT, in the random forest feature selection method with 10-fold cross-validation (Fig. 2). The ML models using hemoglobin A1c and FLI (Model 1) had better abilities for prediction of the development of diabetes mellitus with high AUCs (> 0.85) (Table 2) and high accuracy (> 0.96) (Table 3) than did those using fasting glucose and FLI (Model 2). Although the predictive abilities were not significantly different among the ML models in both Model 1 and Model 2, AUC in logistic regression analysis was the highest. On the other hand, the predictive ability using hemoglobin A1c and components of FLI (Model 3) was significantly higher in logistic regression model than those in models of naïve Bayes and extra gradient boosting. Similarly, the predictive ability using fasting glucose and components of FLI (Model 4) was significantly higher in logistic regression model than that in naïve Bayes model. Taken together, ML models using a combination of hemoglobin A1c and FLI could easily and accurately predict the development of diabetes mellitus in logistic regression model among various ML models.

Several studies have been dedicated to predicting diabetes mellitus using ML models [29–35]. Recent two meta-analyses using 23 ML studies [36] and 12 ML studies [37] demonstrated substantial predictive capabilities of contemporary ML algorithms. In addition to fasting glucose and hemoglobin A1c, physical information including BMI and WC, liver function including GGT and ALT and lipid parameters including TG were selected as important features in previous studies [33–35]. However, FLI has not been investigated as a feature for predicting new onset of diabetes mellitus. The present study identified FLI as the third most influential factor for new onset of diabetes mellitus, suggesting that FLI computation holds promise for guiding health interventions in clinical settings and health checkups for the development of diabetes mellitus as well as the prediction of steatotic liver disease.

In previous studies using traditionally statistical methods, FLI was reported as a significant risk factor for the development of diabetes mellitus [15,38-43]. An association between a high FLI level (≥ 60) as diagnosis of NAFLD and incidence of diabetes mellitus was shown in logistic regression analyses [38–40]. Furthermore, several studies also showed that a high FLI level (≥ 60) was associated with the cumulative incidence of diabetes mellitus in Cox proportional regression analyses using a relatively small number of subjects

(n = 1142∼1922) [41–43]. We previously showed that hazard risk of the development of diabetes mellitus continuously increased with a higher FLI at baseline in both men and women in multivariable Cox proportional hazard models with a restricted cubic spline in 12,290 Japanese subjects (men/women: 7935/4365) [15]. Since there have been no previous studies focused on predictive ability for the development of diabetes mellitus in ML models using FLI as a candidate feature, the present study aimed to fill this research gap by evaluating the predictive performance of various ML models and identifying the most effective predictors.

Various ML models have been widely used as the best-performing algorithms for the development of diabetes mellitus [29–32]. In the present study, prediction capacity in AUCs and accuracies for new onset of diabetes mellitus were comparable among the four ML models including logistic regression, naïve Bayes, extreme gradient boosting and artificial neural network (Table 2). On the other hand, sensitivity was lower in logistic regression model than in other ML models, though specificity in logistic regression model was the highest among ML models (Table 3). Therefore, ML models including naïve Bayes, extreme gradient boosting and artificial neural network using hemoglobin A1c, fasting glucose and FLI without consideration of multicollinearity as the selected features may be useful for identifying individuals at risk for the development of diabetes mellitus. The further development of multiple ML models may be required to achieve more practical predictions for new onset of diabetes mellitus.

The present study has some limitations. First, there were some missing values in 32 of the 58 candidates in the discovery study and in 23 of the 58 candidates in the modeling study. The missing values in parameters with normal distributions and those with skewed distributions were complemented by mean and median values, respectively. Second, several important features including diet, physical exercise and stress were not investigated in the present study. Third, since only Japanese people were enrolled, the results obtained in the present study might not be applicable to other races. Finally, the enrolled subjects had a yearly health check-up at a single urban clinic, the possibility of sample selection bias cannot be ruled out.

In conclusion, ML models incorporating hemoglobin A1c and FLI present an accurate and straightforward approach to predict the development of diabetes mellitus. The calculation of FLI by using non-invasive factors including BMI, WC, GGT and TG would be useful for health guidance in clinical environment and health checkups.

Declarations

Ethics approval and consent to participate

The study conformed to the principles outlined in the Declaration of Helsinki and was conducted with the approval of the Ethics Committee of Sapporo Medical University (Number: 30–2–32). Written informed consent was obtained from all of the subjects.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to ethical restrictions but are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Marenao Tanaka: Writing – original draft, Visualization, Formal analysis, Data curation, Conceptualization. **Yukinori Akiyama:** Supervision, Investigation, Formal analysis. **Kazuma Mori:** Resources, Investigation. **Itaru Hosaka:** Resources, Investigation. **Kenichi Kato:** Resources, Investigation. **Keisuke Endo:** Resources, Investigation. **Toshifumi Ogawa:** Resources, Investigation. **Tatsuya Sato:** Resources, Investigation. **Toru Suzuki:** Resources, Investigation. **Toshiyuki Yano:** Resources, Investigation. **Hirofumi Ohnishi:** Supervision, Formal analysis. **Nagisa Hanawa:** Resources, Investigation, Data curation. **Masato Furuhashi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

M.T, Y.A and M.F were supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (21K09181, 22K08313, 23K07993).

Acknowledgements

The authors are grateful to Keita Numata and Takashi Hisasue for data management.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.deman.2023.100191.

References

[1] American Diabetes Association. Standards of care in diabetes-2023 abridged for primary care providers. Clin Diabetes 2022;41:4–31.
[2] Araki E, Goto A, Kondo T, et al. Japanese clinical practice guideline for diabetes 2019. Diabetol Int 2020;11:165–223.
[3] Cosentino F, Grant PJ, Aboyans V, et al. 2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. Eur Heart J 2020;41:255–323.
[4] Ahmad E, Lim S, Lamptey R, et al. Type 2 diabetes. Lancet 2022;400:1803–20.
[5] Ogurtsova K, da Rocha Fernandes JD, Huang Y, et al. IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract 2017;128:40–50.
[6] Kunutsor SK, Apekey TA, Walley J. Liver aminotransferases and risk of incident type 2 diabetes: a systematic review and meta-analysis. Am J Epidemiol 2013;178:159–71.
[7] Kunutsor SK, Abbasi A, Adler AI. Gamma-glutamyl transferase and risk of type II diabetes: an updated systematic review and dose-response meta-analysis. Ann Epidemiol 2014;24:809–16.
[8] Miyamori D, Tanaka M, Furuhashi M, et al. Prediction of new onset of diabetes mellitus during a 10-year period by using a combination of levels of alanine aminotransferase and gamma-glutamyl transferase. Endocr J 2021;68:1391–402.
[9] Stefan N, Cusi K. A global view of the interplay between non-alcoholic fatty liver disease and diabetes. Lancet Diabet Endocrinol 2022;10:284–96.
[10] Eslam M, Newsome PN, Sarin SK, et al. A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. J Hepatol 2020;73:202–9.
[11] Mendez-Sanchez N, Bugianesi E, Gish RG, et al. Global multi-stakeholder endorsement of the MAFLD definition. Lancet Gastroenterol Hepatol 2022;7:388–90.
[12] Rinella ME, Lazarus JV, Ratziu V, et al. A multi-society Delphi consensus statement on new fatty liver disease nomenclature. J Hepatol 2023.
[13] Bedogni G, Bellentani S, Miglioli L, et al. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006;6:33.
[14] Takahashi S, Tanaka M, Higashiura Y, et al. Prediction and validation of nonalcoholic fatty liver disease by fatty liver index in a Japanese population. Endocr J 2022;69:463–71.
[15] Higashiura Y, Furuhashi M, Tanaka M, et al. High level of fatty liver index predicts new onset of diabetes mellitus during a 10-year period in healthy subjects. Sci Rep 2021;11:12830.
[16] Takahashi S, Tanaka M, Furuhashi M, et al. Fatty liver index is independently associated with deterioration of renal function during a 10-year period in healthy subjects. Sci Rep 2021;11:8606.
[17] Higashiura Y, Furuhashi M, Tanaka M, et al. Elevated fatty liver index is independently associated with new onset of hypertension during a 10-year period in both male and female subjects. J Am Heart Assoc 2021;10:e021430.

[18] Furuhashi M, Muranaka A, Yuda S, et al. Independent association of fatty liver index with left ventricular diastolic dysfunction in subjects without medication. Am J Cardiol 2021;158:139–46.

[19] Mori K, Tanaka M, Higashiura Y, et al. High fatty liver index is an independent predictor of ischemic heart disease during a 10-year period in a Japanese population. Hepatol Res 2022;52:687–98.

[20] Sundstrom J, Schon TB. Machine learning in risk prediction. Hypertension 2020;75:1165–6.

[21] Cutler DR, Edwards Jr. TC, Beard KH, et al. Random forests for classification in ecology. Ecology 2007;88:2783–92.

[22] Behnoush AH, Khalaji A, Rezaee M, et al. Machine learning-based prediction of 1-year mortality in hypertensive patients undergoing coronary revascularization surgery. Clin Cardiol 2023;46:269–78.

[23] Sampson M, Ling C, Sun Q, et al. A new equation for calculation of low-density lipoprotein cholesterol in patients with normolipidemia and/or hypertriglyceridemia. JAMA Cardiol 2020;5:540–8.

[24] Sampson M, Wolska A, Warnick R, et al. A new equation based on the standard lipid panel for calculating small dense low-density lipoprotein-cholesterol and its use as a risk-enhancer test. Clin Chem 2021;67:987–97.

[25] Matsuo S, Imai E, Horio M, et al. Revised equations for estimated GFR from serum creatinine in Japan. Am J Kidney Dis 2009;53:982–92.

[26] Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317–25.

[27] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.

[28] Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. Bone Marrow Transpl 2013;48:452–8.

[29] Gautier T, Ziegler LB, Gerber MS, et al. Artificial intelligence and diabetes technology: a review. Metabolism 2021;124:154872.

[30] Fregoso-Aparicio L, Noguez J, Montesinos L, et al. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr 2021;13:148.

[31] Afsaneh E, Sharifdini A, Ghazzaghi H, et al. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. Diabetol Metab Syndr 2022;14:196.

[32] Mistry S, Riches NO, Gouripeddi R, et al. Environmental exposures in machine learning and data mining approaches to diabetes etiology: a scoping review. Artif Intell Med 2023;135:102461.

[33] Deberneh HM, Kim I. Prediction of type 2 diabetes based on machine learning algorithm. Int J Environ Res Public Health 2021;18.

[34] Kibria HB, Nahiduzzaman M, Goni MOF, et al. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. Sensors (Basel) 2022;22.

[35] Shin J, Kim J, Lee C, et al. Development of various diabetes prediction models using machine learning techniques. Diabetes Metab J 2022;46:650–7.

[36] Silva K, Lee WK, Forbes A, et al. Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis. Int J Med Inform 2020;143:104268.

[37] Kodama S, Fujihara K, Horikawa C, et al. Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: a meta-analysis. J Diabetes Investig 2022;13:900–8.

[38] Balkau B, Lange C, Vol S, et al. Nine-year incident diabetes is predicted by fatty liver indices: the French D.E.S.I.R. study. BMC Gastroenterol 2010;10:56.

[39] Jung CH, Lee WJ, Hwang JY, et al. Assessment of the fatty liver index as an indicator of hepatic steatosis for predicting incident diabetes independently of insulin resistance in a Korean population. Diabet Med 2013;30:428–35.

[40] Yadav D, Choi E, Ahn SV, et al. Fatty liver index as a simple predictor of incident diabetes from the KoGES-ARIRANG study. Medicine (Baltimore) 2016;95:e4447.

[41] Jager S, Jacobs S, Kroger J, et al. Association between the fatty liver index and risk of type 2 diabetes in the EPIC-potsdam study. PLoS One 2015;10:e0124749.

[42] Franch-Nadal J, Caballeria L, Mata-Cases M, et al. Fatty liver index is a predictor of incident diabetes in patients with prediabetes: the PREDAPS study. PLoS One 2018;13:e0198327.

[43] Olubamwo OO, Virtanen JK, Pihlajamaki J, et al. Association of fatty liver index with risk of incident type 2 diabetes by metabolic syndrome status in an Eastern Finland male cohort: a prospective study. BMJ Open 2019;9:e026949.