SUPPLEMENTARY MATERIAL

Supplementary Description of the Sstudy Ppopulation

Participants in this study were derived from the Henan Rural Cohort. The Henan Rural Cohort study has been registered at the Chinese Clinical Trial Register. (Trial registration: ChiCTR-OOC-15006699. Registered 6 July 2015 -Retrospectively registered). The baseline study was conducted in five rural regions in China (Suiping County, Yima City, Tongxu County, Yuzhou City and Xinxiang County) from 2015 to 2017. Briefly, a total number of 39,259 participants aged 18–79 were finally recruited and investigated. Standardized questionnaires that contained a range of indices (including demographic, anthropometric, behaviors, etc.) were used to assess the individual conditions. Biochemical indexes such as fasting blood glucose (FBG) were obtained by physical examinations in standardized processes. After excluding individuals with missing data on family history of T2DM (n = 272), dietary habits (n = 76), physical examination variables (n = 335), cancer (n = 61), kidney failure (n = 4) and air pollutants (n = 253), 38,258 individuals were finally included in this analysis. The flow chart of the data processing procedure is shown in Supplementary Figure S1. This prospective cohort study was sanctioned by the Zhengzhou University Life Science Ethics Committee, and the study was conducted in line with the 1975 Declaration of Helsinki. Before this study, written informed consents were obtained from all participants.

Supplementary Description of the Assessment of Exposures, Outcome and Covariates

According to previous studies of Chen et al., the prediction of individual concentrations of air pollutants was mainly divided into three parts: Estimating spatiotemporal distribution of PM_1 concentrations in China with satellite remote sensing, meteorology, and land use information; estimating $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information; Spatiotemporal patterns of PM_{10} concentrations over China during 2005–2016. The formulas of these studies are displayed here for further researches:

$$\begin{aligned} PM_1 = {} & AODc \times province + s(TEMP) \times province + s(RH) \times province + s(WS) \times province + s(BP) + firesmoke \times province \\ & + NDVI \times province + Forest\_cover + Urban\_cover + Water\_areas + month + Dayofweek + \log(elev) \end{aligned}$$

$$PM_{2.5or10ij} = AOD_{ij} + TEMP_{ij} + RH_{ij} + BP_{ij} + WS_{ij} + NDVI_{ij} + Urban\_cover_{ij} + doy_i + \log(elev_j)$$

$$NO_{2ij} = OMI_{ij} + TEMP_{ij} + BP_{ij} + RH_{ij} + WS_{ij} + NDVI_{ij} + Urban\_cover_{ij} + doy_i + \log(elev_j)$$

where $PM_{2.5\ or\ 10ij}$ represented the $PM_{2.5}$ or PM_{10} on day i at fixed station j; NO_{2ij} represented the NO_2 on day i at fixed station j; $AODc$ or AOD_{ij} exhibited the combined AOD; OMI_{ij} represented the satellite-derived OMI value; *province* represented the fixed station located in province; *TEMP*, *RH*, *BP* and *WS* indicated mean temperature, relative humidity, barometric pressure and wind speed on day i, respectively; *NDVI* represented the monthly average NDVI value at fixed station j; *firesmoke* represented the count of fire smoke spots; *Forest_cover* represented the percentage of forest cover (3-km radius buffer); *Water_areas* represented the percentage of water areas (10-km radius buffer); *Urban_cover* showed the percentage of urban cover with a buffer radius of 10 km around fixed station j; *doy* represented the day of the year; $\log(elev_j)$ meant the log transformed elevation.

According to the recommended criteria of WHO, the definitions of T2DM are listed as follows (type 1 diabetes mellitus, gestational diabetes mellitus and diabetes resulting from other causes are excluded): (1) FBG ≥ 7.0 mmol/L; (2) T2DM patient diagnosed by doctors previously and used anti-glycemic drugs or insulin in the past two weeks; FBG concentrations was tested by the glucose oxidative method using ROCHE Cobas C501 automatic biochemical analyzer (GOD-PAP, Basel, Switzerland).

As for the covariates in this analysis, body mass index (BMI) was calculated as body weight (measured with the participants in light clothing and shoes off by OMRON V. BODY HBF-371, Japan) divided by the square of height (measured to nearest 0.1 cm with shoes off). Waist-to-hip ratio (WHR) was calculated as waist circumference (measured at the level of 1.0 cm above the navel) divided by hip circumference (measured at the maximal level of the hip). The pulse pressure and heart rate were measured and calculated by electronic sphygmomanometers (OMRON HEM-A, Japan) according to the strict operating procedures. The physical activity condition, vegetables and fruits intake, age, gender and family history of T2DM were obtained using standardized questionnaires by trained investigators. Detailed illustration was previously published.

Supplementary Description of the Model Development

Before constructing the model, candidate variables must be determined among more than 100 variables in the data set. As for the specificity of different datasets, Zhang et al. had constructed a nonlaboratory-based risk assessment model for T2DM screening in the Chinese rural population. In this study, we determined the 20 variables selected previously and the air pollutants exposure-related variable as the candidate variables. After that, the univariate Logistic regression was employed to select the variables that were statistically significant. Considering the collinearity of variables, the collinearity diagnosis was employed to eliminate the features with high VIF (> 10). Residual variables were considered predictors and then used to develop the machine learning-model.

As an iterative algorithm, the kernel idea of Gradient Boosting Machine (GBM) is to train different classifiers (weak classifiers) for the same training set and group these weak classifiers to form a more powerful ultimate classifier (strong classifier). According to previous research of Zhang et al., GBM performed best when characterizing T2DM risk by machine learning classifiers, not only for non-laboratory predictors but for laboratory features. Consequently, GBM was applied to model construction with selected variables in the analysis.

When using machine learning classifiers in the field of health prevention, the interpretation of model output was challenging as for the black-box principles. To explain the effect of air pollutants in T2DM risk assessment models, SHAP was employed to show the contribution of predictors as an additive feature attribution method. Each predictor was considered as a “contributor” and then assigned a SHAP value to represent its responsibility for the change in T2DM prevalence. Larger numerical SHAP values indicated greater contributions and the positive direction demonstrated promotion in the prediction. SHAP plots in this analysis were obtained by Jupyter Notebook with a Python 3 kernel. The summary plot was used to show the SHAP ranking of predictors, and the dependency plot showed how the air pollutants depended on other features in the model.

Supplementary Description of Statistical Analysis

(1) detailed description of the QGS formula

$$QGS = \left(\beta_{PM_{2.5}} \times PM_{2.5} + \beta_{PM_{10}} \times PM_{10}\right) \times S_1 + \left(\beta_{PM_1} \times PM_1 + \beta_{NO_2} \times NO_2\right) \times S_2$$

Where: β is the coefficient of the quantile g-computation, which is also called scaled effect size in the Supplementary Table S3; $PM_{2.5}$, PM_{10}, PM_1 and NO_2 are the concentrations of four air pollutants; S is the sum of positive/negative coefficients in the computation.

(2) description of the ambient air pollution score

It’s a simple linear equation weighted by the multivariable-adjusted risk estimates (β coefficients) on T2DM in the present analysis.

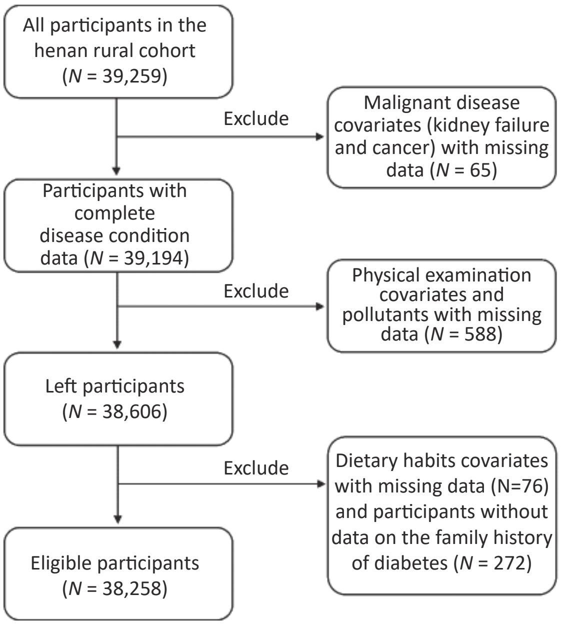
$$APS = \left(\beta_{PM_{2.5}} \times PM_{2.5} + \beta_{PM_1} \times PM_1 + \beta_{PM_{10}} \times PM_{10} + \beta_{NO_2} \times NO_2\right) \times \left(4/\sum \beta\right)$$

Where: β is the coefficient of the univariate logistic regression; PM_{10}, PM_1 and NO_2 are the concentrations of four air pollutants.

(3) description of the Principal Component Analysis (PCA)

PCA was used in this analysis to extract the principal component of four air pollutants, KMO and Bartlett test were used to validate the applicability of this method. A linear model was fitted to calculate the score of ambient air pollution with the variance relative contribution of each component. The score is called AAP score in this analysis.

$$AAP\ Score = (V_1\sum\nolimits_{i=1}^{n} x_i\beta_i + V_2\sum\nolimits_{i=1}^{n} x_i\beta_i)/(V_1 + V_2)$$


All participants in the henan rural cohort (*N* = 39,259)
Exclude
Malignant disease covariates (kidney failure and cancer) with missing data (*N* = 65)
Participants with complete disease condition data (*N* = 39,194)
Exclude
Physical examination covariates and pollutants with missing data (*N* = 588)
Left participants (*N* = 38,606)
Exclude
Dietary habits covariates with missing data (N=76) and participants without data on the family history of diabetes (*N* = 272)
Eligible participants (*N* = 38,258)

Supplementary Figure S1. The data processing flow chart of this study.

Supplementary Table S1. Characteristics of the traditional T2DM predictors

Characteristics	Non-T2DM (*n* = 34,694)	T2DM (*n* = 3,564)	*P* value*
Age, mean ± SD, years	55.13 ± 12.34	60.47 ± 9.25	< 0.001
Men (*n*, %)	13,691 (39.46)	1,340 (37.60)	0.031
Body mass index, mean ± SD, kg/m^2	24.67 ± 3.50	26.13 ± 3.66	< 0.001
Waist-to-hip ratio, mean ± SD	0.88 ± 0.07	0.93 ± 0.07	< 0.001
Heart rate, mean ± SD, beats/minute	75.34 ± 10.98	79.62 ± 12.21	< 0.001
Pulse pressure, mean ± SD, mmHg	47.69 ± 12.83	53.36 ± 14.23	< 0.001
More vegetable and fruit intake (yes, *n*, %)	14,787 (42.62)	1,269 (35.61)	< 0.001
Physical activity (*n*, %)			
Low	10,915 (31.46)	1,389 (38.97)	< 0.001
Moderate	13,296 (38.32)	1,268 (35.58)	< 0.001
High	10,483 (30.22)	907 (25.45)	< 0.001
Family history of T2DM (yes, *n*, %)	1,229 (3.54)	352 (9.88)	< 0.001

Note. SD indicated standard error; *Student's *t*-test was used to compare the mean difference of continuous variables; Chi-square test was used to test the distributions of categorical variables.

Supplementary Table S2. Characteristics of the ambient air pollutants

Characteristics	Non-T2DM (*n* = 34,694)	T2DM (*n* = 3,564)	*P* value[*]
NO_2, mean ± SD, μg/m^3	39.80 ± 3.61	40.66 ± 3.53	< 0.001
PM_1, mean ± SD, μg/m^3	57.41 ± 2.67	57.81 ± 2.67	< 0.001
$PM_{2.5}$, mean ± SD, μg/m^3	73.36 ± 2.58	73.95 ± 2.50	< 0.001
PM_{10}, mean ± SD, μg/m^3	132.32 ± 5.85	133.76 ± 5.61	< 0.001

Note. [*]Student's *t*-test was used to compare the mean difference of continuous variables.

Supplementary Table S3. Coefficients of the quantile g-computation in this study

Items	Scaled effect size (***β***)[*]
Positive direction	
$PM_{2.5}$	0.709
PM_{10}	0.291
Sum of positive coefficients (S1)	0.407
Negative direction	
NO_2	0.501
PM_1	0.439
Sum of negative coefficients (S2)	−0.194

Note. [*]Adjusted for the age, gender and physical activity.

Supplementary Table S4. Associations (*OR*s and 95% *CI*) of the mixture of ambient air pollutants with T2DM

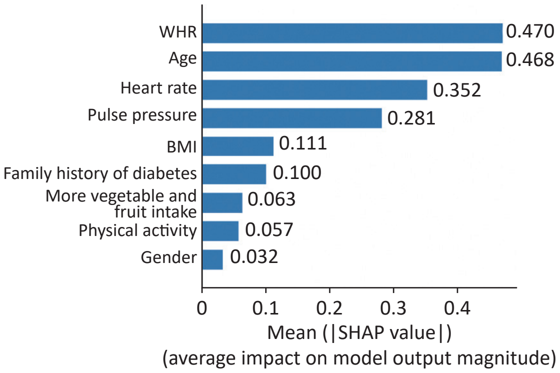
Variables	Model 1	Model 2	Model 3
AAP score			
Each unit change	1.03 (1.02, 1.03)	1.02 (1.02, 1.02)	1.02 (1.01, 1.02)
T1[*]	1.00 (Ref.)	1.00 (Ref.)	1.00 (Ref.)
T2	1.54 (1.41, 1.68)	1.36 (1.23, 1.49)	1.29 (1.17, 1.42)
T3	1.81 (1.66, 1.98)	1.55 (1.41, 1.70)	1.45 (1.31, 1.60)
APS			
Each unit change	1.02 (1.02, 1.02)	1.01 (1.01, 1.02)	1.01 (1.01, 1.01)
T1[*]	1.00	1.00	1.00
T2	1.52 (1.39, 1.66)	1.33 (1.21, 1.46)	1.26 (1.15, 1.39)
T3	1.83 (1.68, 2.00)	1.57 (1.43, 1.72)	1.47 (1.33, 1.62)

Note. AAP score, the ambient air pollution score; APS, the air pollution score; T1-T3 were the tertiles of the AAP score and APS; Model 1 was the crude model; Model 2 adjusted for the age, gender, BMI, waist-to-hip ratio, pulse pressure and heart rate; Model 3 was further adjusted for more vegetable and fruit intake, physical activity and family history of diabetes. [*]Different groups were divided by the tertiles of the AAP score and APS.

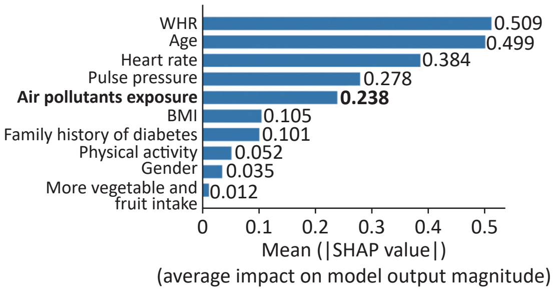
Supplementary Table S5. Performance metrics of different machine learning classifiers

	AUC (traditional)	BS (traditional)	AUC (traditional & air pollutants exposure*)	BS (traditional & air pollutants exposure*)
GBM	0.764	0.079	0.787	0.076
RF	0.761	0.082	0.762	0.082
ANN	0.712	0.086	0.732	0.081

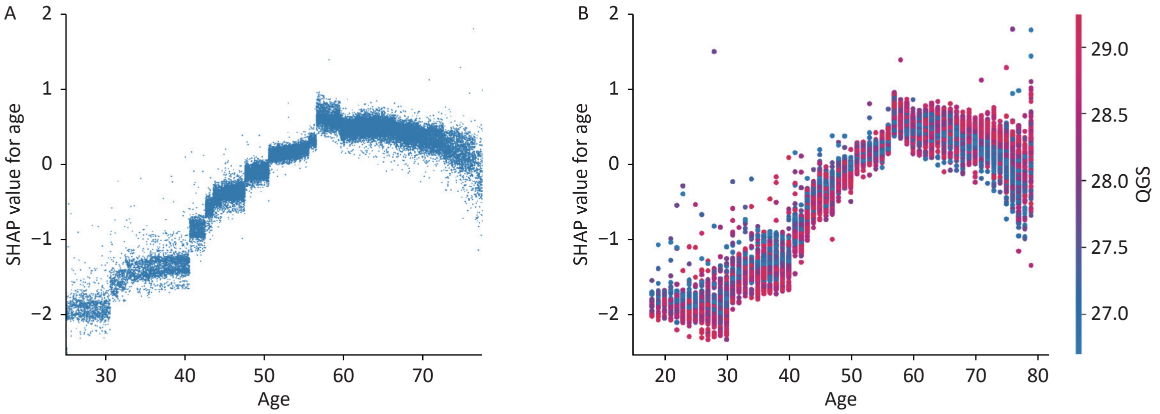
Note. *Calculated by the quantile g-computation. BS, brier score.



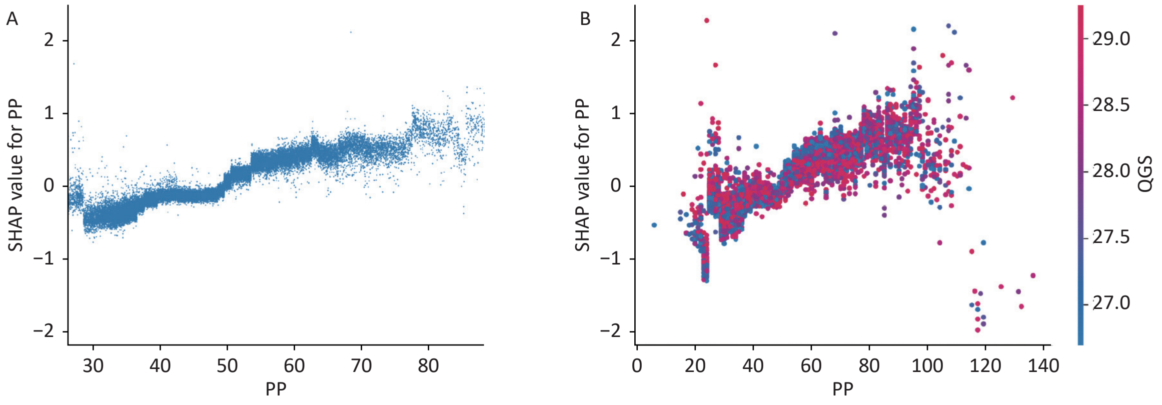
Supplementary Figure S2. Feature importance of GBM by summing of SHAP value magnitudes (the bar plot of traditional predictors). SHAP, SHapely additive exPlanations. GBM, Gradient Boosting Machine.



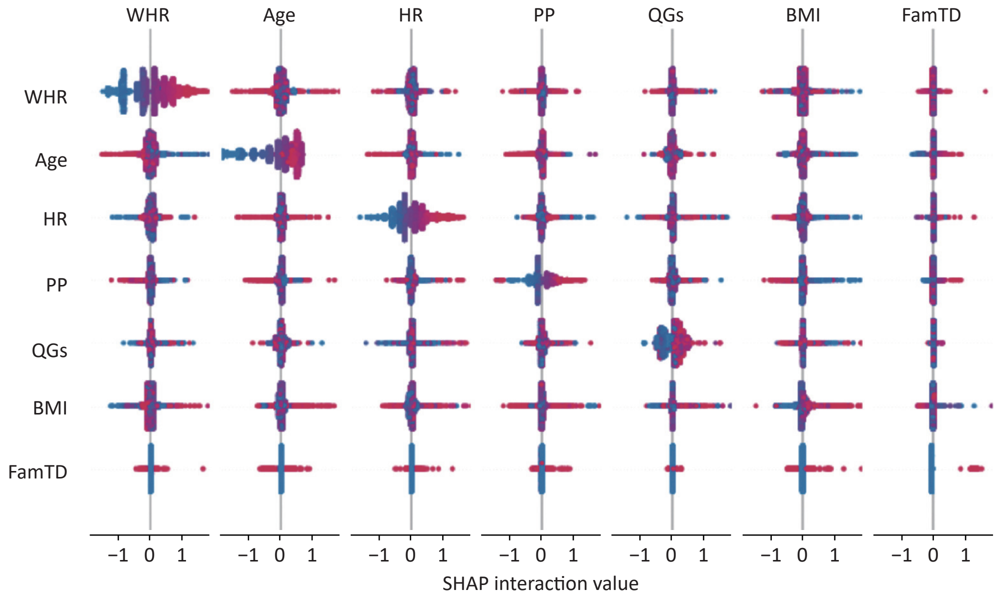
Supplementary Figure S3. Feature importance of GBM by summing of SHAP value magnitudes (the bar plot of predictors considering air pollutants).



Supplementary Figure S4. Main and interaction effect of important features of GBM. (A) SHAP dependence plot for age on T2DM. (B) SHAP dependence plot of age with interaction of air pollutants exposure.SHAP, SHapely additive exPlanations. GBM, Gradient Boosting Machine.



Supplementary Figure S5. Main and interaction effect of the pulse pressure (PP) with QGS (A) SHAP dependence plot for PP on T2DM (B) SHAP dependence plot of PP with interaction of QGS.SHAP, SHapely additive exPlanations.



Supplementary Figure S6. Interaction effect of all features of GBM in the traditional & APE model. FamTD indicates family history of diabetes. Different colors are linked with levels of variables, and red means higher level, blue means lower level. GBM, Gradient Boosting Machine.