

Supplemental information

FetalADM: a double-layer ensemble model based on SMOTE and RF-RFE for fetal trisomy 21, 18, and 13 detections

Xiaohan Sun ^{a,1}, Jianjiang Zhu ^{b,1}, Xuequn Mao ^b, Yousheng Yan^c, Limei Xu ^b, Wen Zeng ^b, Hong Qi ^{b*} and Jianbo Lu^{d*}

^a School of Sports and Health, Tianjin University of Sport, Tianjin, 301617, China

^b Prenatal Diagnosis Center, Haidian District Maternal and Child Health Care Hospital, Beijing 100080, China

^c Prenatal Diagnostic Center, Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100026, China

^d National Human Genetics Resource Center, National Research Institute for Family Planning, 100081, Beijing, China

¹ Xiaohan Sun and Jianjiang Zhu should be considered joint first author.

*Corresponding author

Jianbo Lu: Email: jblu@lsec.cc.ac.cn

Hong Qi: Email: qihong_2009@sina.com

SMOTE_KM algorithm

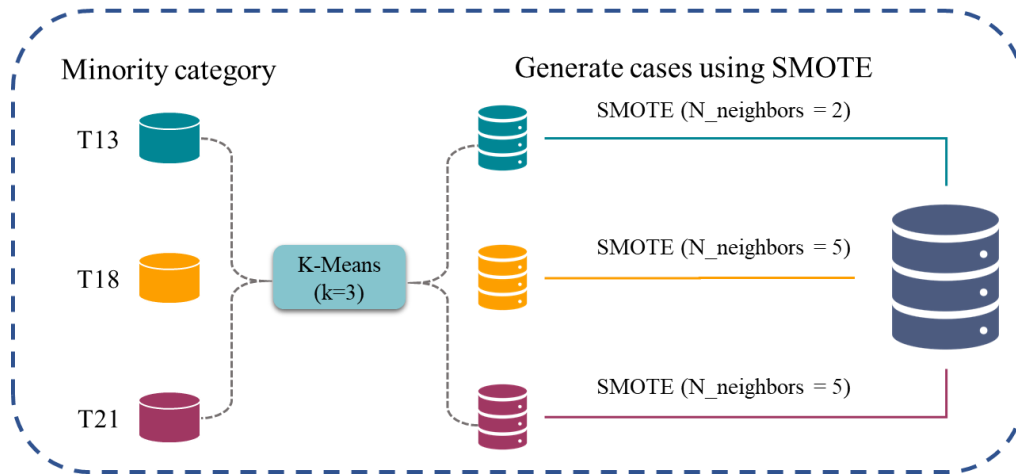


Figure S1. The process of the SMOTE_KM algorithm. The process of the algorithm as follows: firstly, we used K-Means clustering to the data in each minority category (T13, T18 and T21) to divide the samples into k clusters. The value of k in K-Means was determined based on the sample size of fetal trisomy 13 (T13), which is the smallest minority class. Specifically, we applied K-Means to each minority category (T13, T18, and T21) and divided the samples into 3 clusters. After clustering, SMOTE is applied to generate synthetic samples. For each sample which x_i is contained minority class, a sample $\bar{x}_j$ is randomly selected from its n nearest neighbors. Then, a point on the line between the sample x_i and $\bar{x}_j$ is randomly selected as the newly synthesized minority class. In the SMOTE step, the neighbor size was set to the default value ($N_neighbors = 5$), except for T13. Due to the extremely small sample size of T13 ($n=3$), the number of neighbors was set to 2 to satisfy computational constraints and ensure successful synthetic sample generation.

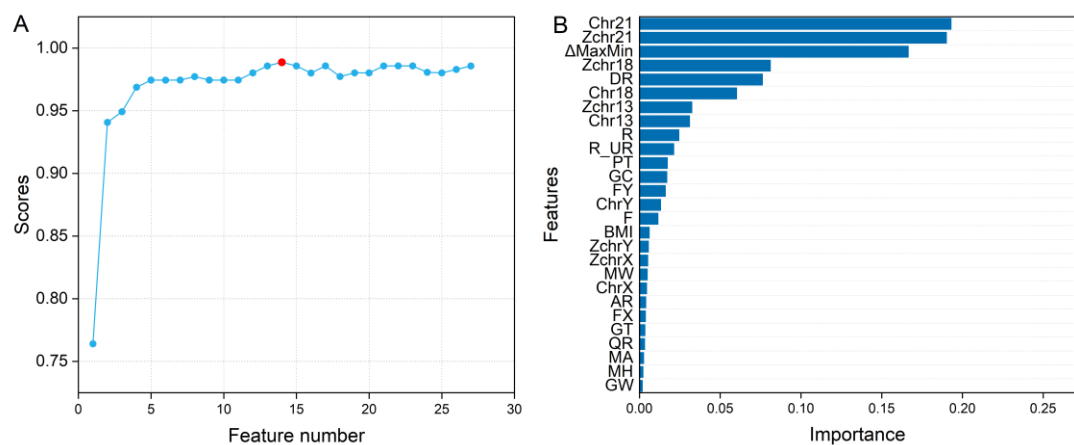


Figure S2. Results of feature selection. (A) The x-coordinate represents the number of features and the y-coordinate represents the accuracy. In this algorithm, the cross-validation (cv) is implemented using StratifiedKFold with n_splits set to 10 and shuffle enabled (cv = StratifiedKFold (n_splits=10, shuffle=True)). The red circle means the number of features is 14. (B) The histogram of feature importance ranking.

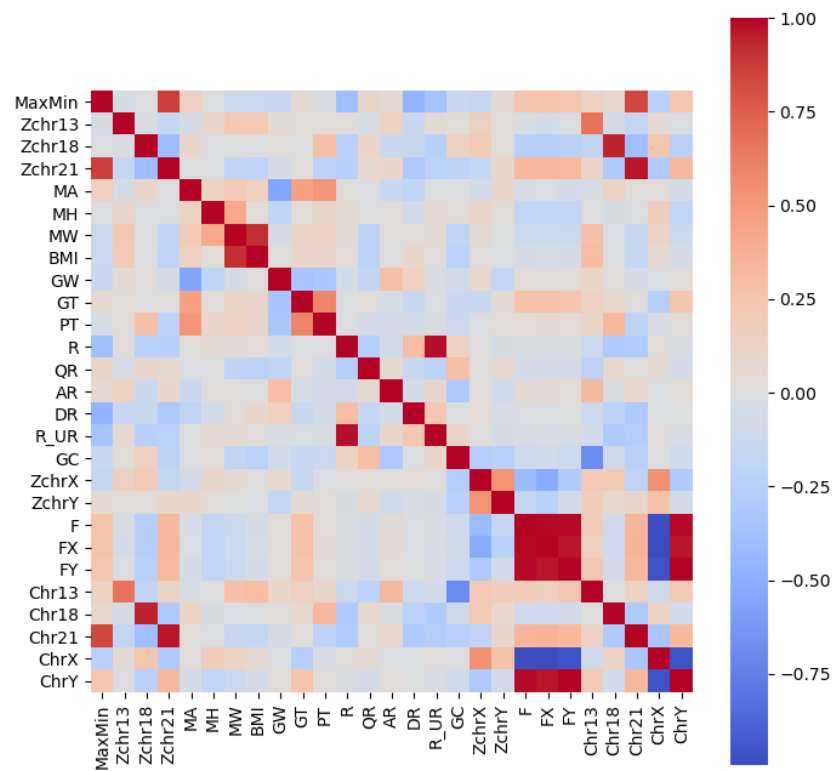


Figure S3. The heatmap illustrating the relationships among the 27 features.

MaxMin represents our designed feature ΔMaxMin .

Table S1. The descriptions of 26 features.

Features	Descriptions
MA	Maternal age
MH	Maternal height
MW	Maternal weight
BMI	Body Mass Index
GW	Gestational week
GT	Gravidity times
PT	Parity times
R	Sequenced reads
DR	Duplicate rate
R_UR	Real unique reads
GC	GC depth
QR	Q30 rate
AR	Alignment rate
Zchr13	Z-score of chromosome 13
Zchr18	Z-score of chromosome 18
Zchr21	Z-score of chromosome 21
ZchrX	Z-score of chromosome Y
ZchrY	Z-score of chromosome Y
F	Fetal fraction in maternal plasma
FX	Fraction of chromosome X in maternal plasma
FY	Fraction of chromosome Y in maternal plasma
Chr13	Ratio of chromosome 13 fragments
Chr18	Ratio of chromosome 18 fragments
Chr21	Ratio of chromosome 21 fragments
ChrX	Ratio of chromosome X fragments
ChrY	Ratio of chromosome Y fragments
MaxMin	We designed

Table S2. Comparison of models with and without SMOTE_KM on model performances.

	Methods	ACC	PRE	REC	F1	MCC
With SMOTE_KM	XGBoost	0.972	0.974	0.972	0.972	0.955
	RF	0.972	0.974	0.972	0.972	0.955
	SVM	0.944	0.948	0.944	0.945	0.908
Without SMOTE_KM	XGBoost	0.963	0.964	0.963	0.962	0.878
	RF	0.930	0.912	0.93	0.915	0.767
	SVM	0.884	0.828	0.884	0.847	0.580

The mean and standard deviation (the values in parentheses) of the metrics were obtained by averaging the results across five-fold cross-validation.