Machine/Deep Learning-based Approaches to Predict Overweight or Obesity in Chinese Preschool-Aged Children

Qiong Wang
China-Japan Friendship Hospital
Min Yang
China-Japan Friendship Hospital
Bo Pang
China-Japan Friendship Hospital
Mei Xue
China-Japan Friendship Hospital
Yicheng Zhang
China-Japan Friendship Hospital
Xiangling Deng
China-Japan Friendship Hospital
Zhixin Zhang
China-Japan Friendship Hospital
Wenquan Niu (niuwenquan_shcn@163.com)
China Japan Friendship Institute of Clinical Medicine Research https://orcid.org/0000-0003-1715-3372

Research Article

Keywords: Obesity, Overweight, Preschool-aged children, Factor, Machine Learning, Prediction Model

Posted Date: February 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1311362/v1

Abstract

Objectives: We adopted the machine learning and deep learning algorithms to explore risk profiling for overweight and obesity in Chinese preschool-aged children.

Methods: This is a cross-sectional survey, and was conducted between September and December in 2020 at Beijing and Tangshan. Using a stratified cluster random sampling strategy, children 3-6 years of age from 30 kindergartens were enrolled. Data were analyzed using community PyCharm.

Results: A total of 9478 children were eligible for inclusion, 1250 children with overweight or obesity, and they were randomly divided into the training group and testing group in a 6:4 ratio. After the training and testing process, the SVM (accuracy: 0.9457) was ranked as the best algorithm, followed by the GBM (accuracy: 0.9454) as reflected by model accuracy. As reflected by other performance indexes, the GBM had the highest F1 score (0.7748), followed by the SVM with the F1 score at 0.7731. After importance ranking, the top five factors seemed sufficient to obtain descent performance under the GBM algorithm, including age of children, eating speed, number of relatives with obesity, sweet drinking, and paternal education, which was further reinforced by a classical deep learning sequential model.

Conclusions: We identified five factors relating to children and parents that can help differentiate children with overweight or obesity from general children. Further validations are necessary.

Introduction

Globally, childhood overweight/obesity has reached epidemic proportions and represents a major public health concern [1]. From 1975 to 2016, the prevalence of overweight in children and adolescents was increased from 0.7–5.6% in girls and 0.9–7.8% in boys around the world [2, 3]. Latest statistics indicate an estimated 38 million children under five years of age are overweight or obese, and this number skyrockets to approximate 340 million among children and adolescents aged 5 to 19 years [4]. In China, due to the dramatic socio-economic changes and nutritional transitions during the past decades, the prevalence of overweight/obesity was increased from 11.7–25.2% during the period between 1991 and 2011 [5]. Considering that the detrimental impact of childhood obesity can extend into adulthood and precipitate the development of metabolic abnormalities and pulmonary disorders [6], it is of public health importance to curb this inclination for society as a whole by gaining a better understanding of the risk profiles of childhood obesity [7, 8].

Identification and characterization of potential risk factors responsible for the development of overweight or obesity in children have been widely undertaken [9, 10]. For example, reducing calorie intake and increasing physical activity were recommended to children with overweight or obesity [11–13]. As reported previously, our findings indicated that bedtime and eating speed acted synergistically in predicting the risk of overweight and obesity among 1123 preschool-aged children [14]. Further exploration among 7222 preschool-aged children identified four factors, including maternal body mass index (BMI), maternal pre-pregnancy BMI, breastfeeding duration and sleep duration, in significant

association with childhood overweight or obesity [15]. Despite intensive efforts devoted to seek obesity-susceptibility factors in children and adolescents, there is thus far no definite consensus on how many factors and which one(s) actually play the role.

Currently, a major challenge facing the majority of previous studies is to better define the complex relationship between these factors and outcomes [16]. This is because traditional prediction models are limited by lack of inclusion of nonlinear, collinear and interactive effects among factors [17]. In response to this limitation, a wide panel of machine learning and deep learning algorithms are developed as advanced statistical tools to facilitate the characterization of these effects [16, 18]. Machine learning and deep learning methods have been successfully applied for a variety of clinical endpoints in children including, for instance, myopia [19], dental caries [20], depression [21], obesity [16] and so on.

To yield more information, we attempted to adopt the widely-used supervised machine learning and deep learning algorithms to explore the risk profiling for overweight and obesity in a large survey of preschool-aged children from 30 kindergartens located in Beijing and Tangshan. Specifically, under the best algorithm selected, the importance of factors under evaluation was ranked first. Next, the optimal number of important factors and the best algorithm in prediction of childhood overweight/obesity were ascertained. Finally, a visualization tool was developed for wide application.

Methods

Study design

This survey is cross-sectional in nature, and was conducted during the period between September and December in 2020 at Beijing and Tangshan (Hebei province), China. The implementation of this survey received approval from the Ethics Committee of China-Japan Friendship Hospital, and was in compliance with the principles of the Declaration of Helsinki. Signed informed consent was obtained from the parents of assessable children who participated in the present study.

Study participants

A stratified cluster random sampling strategy was used to collect information from preschool-aged kindergarten children in Beijing and Tangshan. In detail, four districts out of 16 districts in Beijing and two districts out of seven districts in Tangshan were selected. Within each district, five kindergartens were randomly selected, and so a total of 30 kindergartens entered into this survey. Children from these 30 kindergartens formed the study participants, except those who were diagnosed to have major illnesses, which include but not limit to chronic kidney disease, hypothyroidism, or congenital heart disease.

Data collection and quality control

Data were collected by circulating our self-designed questionnaires, which were a priori found to have reliability coefficient alpha over 0.85, to the parents or guardians of a total of 10441 children from selected kindergartens. The questionnaires were filled in online via the tool termed as the WenJuanXing

(https://www.wenjuan.com/), and finally 10230 questionnaires were returned with a response rate of 98%. Data from completed questionnaires were downloaded in the form of Excel from this website.

Information from questionnaires was collected from both children and their parents. From children, sex, region, date of birth, time spent on outdoor activities at workdays and weekends, weekly intake frequency of fast food and night meals, picky eating, birthweight, birth height, gestational age, delivery mode, twin birth, birth order, breastfeeding duration and solid food introduction age were recorded. Thereof, weight (to the nearest 0.1 kg) and height (to the nearest 0.1 cm) were measured by trained healthcare physicians

From parents, self-reported data included age, height, gestational diabetic mellitus, education, family income, perinatal clinical history (delivery mode, birth weight, and birth height), duration of breast-feeding, and self-rated patience to children.

Kindergarten teachers were responsible for sending electronic questionnaires online to the parents or guardians of all participant children. Data exported from electronic questionnaires to a Microsoft Office Excel^{TM} spreadsheet were strictly checked by trained staff. In case of missing or uncertain records, parents or guardians were contacted by phones for the sake of accuracy.

Overweight and obesity definition

Several official definitions are available for the definition of childhood overweight and obesity, including the International Obesity Task Force (IOTF) criteria, World Health Origination (WHO) criteria, and Chinese criteria. In this study, we adopted the WHO criteria for wide application. In detail, overweight and obesity are defined based on body mass index (BMI) z-scores at a cutoff of 5 years old under the WHO criteria [22–24]. In children 5 years of age or below, overweight and obesity are defined as the BMI Z-score between 2 and 3 and > 3, respectively. In children over 5 years of age, overweight and obesity are separately defined as the BMI Z-score between 1 and 2 and > 2.

Definitions of baseline characteristics

Time spent on outdoor activities every day was calculated as the sum of time both on workdays × 5 and weekends × 2 divided by 7. Fast foods referred to foods with high energy and low nutrition (e.g., hamburger and French fries), Night meal was defined as eating food within 2 hours before bedtime. Weekly intake frequency was consistent with fast food and night meals, which was classified as every day, often (3-5 times), occasional (1-2 times) or none or occasionally. Picky eating was defined as yes or no. Gestational, birth weight and birth height were recorded. Delivery mode included vaginal delivery and caesarean section.

Parental height was self-reported. Paternal and maternal age at delivery was calculated as the difference between the date of child's birthdate and parent birthdate. Maternal gestational diabetes mellitus diagnosed by doctors from second-class or above hospitals, were recorded. Education was categorized as doctor's degree or above, master's degree, bachelor's degree, and high school degree or below. The relatives in this study referred to the parents, grandparents, and grandparents-in-law of children. Family

income (RMB per year) was categorized as ≥1,000,000, 600,000-1,000,000, 300,000-600,000, 100,000-300,000, and <100,000.

Statistical analyses

Considering that the number of children with obesity was small, overweight and obesity were combined as a single group compared with the non-overweight group (the reference group). Factors with missing data over 30% were removed from the analysis. Missing data were derived using the multiple imputation procedure by the mice package in the R environment (Version 4.1.1). Categorical data are expressed as count (percentage). For continuous data, P values for comparison between children with non-overweight and overweight or obesity were derived by the t test for normally distributed data, the rank-sum test for skewed data, and the χ^2 test for categorical data.

To examine the association of data from children and parents with childhood overweight/obesity, nine machine learning algorithms were employed, including Logistic regression model, decision tree, support vector machine (SVM), random forest, K-nearest neighbor (KNN), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), and naïve Bayes. On the basis of the nine machine learning algorithms, both hard and soft voting classifiers were calculated as an ensemble machine learning method. The performance of machine learning algorithms was assessed using accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUROC). Accuracy is a measurement of how good a model is. Precision is a measurement of how many positive predictions were actual positive observations. Recall is a measure of how many actual positive observations were predicted correctly. F1 score is an ‘average’ of both precision and recall. The importance of each factor under study was calculated using the χ^2 test and ranked in an ascending order.

Additionally, a deep learning algorithm, sequential model was also employed to test this association by using three different optimization algorithms, that is, adaptive moment estimation (Adam), root mean square prop (RMSprop), and stochastic gradient descent (SGD). Model loss and accuracy were used to appraise prediction performance.

Both machine learning and deep learning algorithms were trained on 60% of study children (the training group) and tested on the remaining 40% (the testing group) as an internal validation of the prediction model.

All analyses were done using the community PyCharm (Edition 2018.1 x64) under Windows 10 with the Python (Python Software Foundation) software (Version 3.7.6).

Results

Baseline characteristics

After removing data with incomplete information of interest, a total of 9478 children were eligible for inclusion, and their baseline characteristics by obesity status are presented in Table 1. There are 1250 children with overweight or obesity, which accounts for 13.19% of all study children.

Table 1
The baseline characteristics of study children by overweight/obesity status.

Factors under study		Non-overweight	Overweight or obesity	P
		(n=8228)	(n=1250)	
Baseline factors				
Sex (%)	Boys	4134 (50.2%)	677 (54.2%)	0.011
	Girls	4094 (49.8%)	573 (45.8%)	
Age(month)		54.70 [48.20, 65.40]	66.40 [55.30, 72.70]	<0.001
Lifestyle-related factors				
Night meal (%)	None or occasionally	4801 (58.3%)	762 (61.0%)	0.002
	1-2 times weekly	1872 (22.8%)	299 (23.9%)	
	3-5 times weekly	795 (9.7%)	113 (9.0%)	
	Every day	760 (9.2%)	76 (6.1%)	
Sweet foods (%)	None or occasionally	484 (5.9%)	67 (5.4%)	0.218
	1-2 times weekly	2046 (24.9%)	284 (22.7%)	
	3-5 times weekly	4452 (54.15%)	714 (57.1%)	
	Every day	1246 (15.1%)	185 (14.8%)	
Sweet drinking (%)	None or occasionally	131 (1.6%)	21 (1.7%)	<0.001
	1-2 times weekly	220 (2.7%)	44 (3.5%)	
	3-5 times weekly	1923 (23.4%)	409 (32.7%)	
	Every day	5954 (72.4%)	776 (62.1%)	
Fast foods (%)	None or occasionally	62 (0.8%)	14 (1.1%)	<0.001
	1-2 times weekly	76 (0.9%)	19 (1.5%)	

Continuous data are expressed as mean (standard deviation) in normal distributions and median [interquartile range] in skewed distributions. Categorical data are expressed as count (percentage). For continuous data, P for comparison between children with non-overweight and overweight or obesity was derived by t test for normally distributed data, by rank-sum test for skewed data, and by χ^2 test for categorical data.

Factors under study		Non-overweight	Overweight or obesity	P
		(n=8228)	(n=1250)	
	3-5 times weekly	2582 (31.4%)	464 (37.1%)	
	Every day	5508 (66.9%)	753 (60.2%)	
Eating speed (minute)		18.30 [15.00, 26.70]	16.70 [13.30, 23.30]	<0.001
Outdoor activities (h per day)		1.60 [1.00, 2.30]	1.60 [1.00, 2.30]	0.95
Sitting duration (h per day)		2.00 [1.30, 3.70]	2.30 [1.30, 4.10]	<0.001
Electronic screens (h per day)		1.00 [0.60, 1.60]	1.10 [0.60, 1.90]	<0.001
Sleep duration (h per day)		10.00 [9.30, 10.60]	9.90 [9.00, 10.30]	<0.001
Fall asleep time (h per day)		9.00 [9.00, 10.00]	9.00 [9.00, 10.00]	0.385
Fetal and neonatal factors				
Pregnancy order (median [IQR])		2.00 [1.00, 2.00]	2.00 [1.00, 2.00]	0.117
Delivery order		1.00 [1.00, 2.00]	1.00 [1.00, 2.00]	0.566
Delivery mode (%)	Vaginal delivery	4500 (54.7%)	584 (46.7%)	<0.001
	Cesarean section	3728 (45.3%)	666 (53.3%)	
Birthweight (kg)		3.30 [3.00, 3.60]	3.40 [3.00, 3.75]	<0.001
Birth body length (cm)		50.00 [50.00, 52.00]	51.00 [50.00, 52.00]	<0.001
Bearing age of father		30.60 [27.90, 34.20]	29.95 [27.22, 33.40]	<0.001
Bearing age of mother		29.30 [27.00, 32.70]	28.90 [26.40, 32.00]	<0.001

Continuous data are expressed as mean (standard deviation) in normal distributions and median [interquartile range] in skewed distributions. Categorical data are expressed as count (percentage). For continuous data, P for comparison between children with non-overweight and overweight or obesity was derived by t test for normally distributed data, by rank-sum test for skewed data, and by χ^2 test for categorical data.

Factors under study		Non-overweight	Overweight or obesity	P
		(n=8228)	(n=1250)	
Activities time during pregnancy (h per day)		1.60 [1.00, 2.60]	2.00 [1.00, 3.00]	<0.001
Gestational diabetes (%)	NO	7488 (91.0%)	1126 (90.1%)	0.314
	YES	740 (9.0%)	124 (9.9%)	
Gestational weight gain (kg)		15.00 [10.00, 18.00]	15.00 [10.00, 20.00]	<0.001
Infancy feeding (%)	Breastfeeding	4738 (57.6%)	688 (55.0%)	0.096
	Non-breastfeeding	3490 (42.4%)	562 (45.0%)	
Breastfeeding duration (months)		13.00 [8.00, 18.00]	12.00 [8.00, 18.00]	0.038
Family-related factors				
Number of relatives with obesity	0	5680 (69.0%)	736 (58.9%)	<0.001
	1	1747 (21.2%)	307 (24.6%)	
	2	578 (7.0%)	141 (11.3%)	
	3	157 (1.9%)	51 (4.15%)	
	4	45 (0.5%)	13 (1.0%)	
	5	15 (0.2%)	2 (0.25)	
	6	6 (0.1%)	0 (0.0%)	
Number of relatives with diabetes	0	5693 (69.25)	828 (66.2%)	0.226
	1	2038 (24.8%)	331 (26.5%)	
	2	433 (5.3%)	74 (5.95)	
	3	47 (0.6%)	12 (1.0%)	
	4	6 (0.1%)	2 (0.2%)	
	5	6 (0.1%)	2 (0.2%)	

Continuous data are expressed as mean (standard deviation) in normal distributions and median [interquartile range] in skewed distributions. Categorical data are expressed as count (percentage). For continuous data, P for comparison between children with non-overweight and overweight or obesity was derived by t test for normally distributed data, by rank-sum test for skewed data, and by χ^2 test for categorical data.

Factors under study		Non-overweight	Overweight or obesity	P
		(n=8228)	(n=1250)	
	6	5 (0.1%)	1 (0.1%)	
Number of relatives with asthma	0	7881 (95.8%)	1194 (95.5%)	0.463
	1	327 (4.0%)	50 (4.0%)	
	2	9 (0.1%)	3 (0.2%)	
	3	4 (0.0%)	2 (0.2%)	
	6	7 (0.1%)	1 (0.1%)	
Number of relatives with allergy	0	5850 (71.1%)	884 (70.7%)	0.113
	1	1833 (22.3%)	282 (22.6%)	
	2	447 (5.4%)	61 (4.9%)	
	3	83 (1.0%)	17 (1.4%)	
	4	6 (0.1%)	4 (0.3%)	
	5	4 (0.0%)	0 (0.0%)	
	6	5 (0.1%)	2 (0.2%)	
Paternal education	High school degree or below	2868 (34.9%)	551 (44.1%)	<0.001
	Bachelor’s degree	3925 (47.7%)	535 (42.8%)	
	Master’s degree	1011 (12.3%)	119 (9.5%)	
	Doctor’s degree or above	424 (5.2%)	45 (3.6%)	
Maternal education	High school degree or below	2611 (31.7%)	486 (38.9%)	<0.001
	Bachelor’s degree	4277 (52.0%)	619 (49.5%)	
	Master’s degree	1105 (13.45)	119 (9.5%)	

Continuous data are expressed as mean (standard deviation) in normal distributions and median [interquartile range] in skewed distributions. Categorical data are expressed as count (percentage). For continuous data, P for comparison between children with non-overweight and overweight or obesity was derived by t test for normally distributed data, by rank-sum test for skewed data, and by χ^2 test for categorical data.

Factors under study		Non-overweight	Overweight or obesity	P
		(n=8228)	(n=1250)	
	Doctor's degree or above	235 (2.9%)	26 (2.1%)	
Family income (RMB per year)	<100000 RMB	2741 (33.3%)	474 (37.9%)	0.006
	100000-300000 RMB	3032 (36.8%)	458 (36.6%)	
	300000-600000 RMB	1693 (20.6%)	219 (17.5%)	
	600000-1000000 RMB	487 (5.9%)	60 (4.8%)	
	> 1000000 RMB	275 (3.3%)	39 (3.1)	

Continuous data are expressed as mean (standard deviation) in normal distributions and median [interquartile range] in skewed distributions. Categorical data are expressed as count (percentage). For continuous data, P for comparison between children with non-overweight and overweight or obesity was derived by t test for normally distributed data, by rank-sum test for skewed data, and by χ^2 test for categorical data.

Machine learning models

To select the model with the best performance, children's data were evaluated by nine widely-evaluated machine learning algorithms. After the training and testing process, the SVM (accuracy: 0.9457) was ranked as the best algorithm, followed by the GBM (accuracy: 0.9454) as reflected by model accuracy (Figure 1). Importantly, the performance of both algorithms was superior over that of hard (accuracy: 0.9436) and soft (accuracy: 0.9433) voting classifications.

Besides accuracy, four additional performance indexes were also evaluated under the nine machine learning algorithms. As shown in Figure 2, out of all algorithms, the GBM had the highest F1 score (0.7748), followed by the SVM with the F1 score at 0.7731. Considering that F1 score is the harmonic mean of precision and recall, the GBM was identified as the best machine learning model in this study.

Ranking importance

As this study involved an analysis of 31 factors, it is interesting to know the importance order of these factors. To shed some light, the importance of top ten factors was estimated and ranked, as displayed in Figure 3. Overall, the most important factor was age of children, followed by eating speed, number of relatives with obesity, sweet drinking, paternal education, delivery mode, number of relative with diabetes, night meals, family incoming, and maternal education.

Selection of best factors

To select the minimum number of factors with decent prediction performance, the changes in AUROC, accuracy, and precision with the increasing number of top factors using the GBM algorithm are presented in Table 2. By comparison, the top five factors seemed sufficient enough to obtain descent performance, and they included age of children, eating speed, number of relatives with obesity, sweet drinking, and paternal education.

Table 2
The distributions of areas under the receiver operating curve (AUROC), accuracy and precision with the cumulating number of top ten factors in an ascending order.

Number of top ten factors in rank	AUROC	Accuracy	Precision
1	0.9527	0.9456	0.8258
2	0.9538	0.9456	0.8258
3	0.9539	0.9443	0.8177
4	0.9539	0.9435	0.8225
5	0.9837	0.9443	0.8222
6	0.9836	0.9430	0.8141
7	0.9842	0.9438	0.8276
8	0.9838	0.9427	0.8151
9	0.9835	0.9433	0.8221
10	0.9836	0.9411	0.7968

Confirmation by deep learning model

To examine whether the performance of above top five factors is comparable with all factors under study, a classical deep learning sequential model was used in both training and testing groups (Table 3). To optimize this model, three optimizers were adopted. From both loss and accuracy aspects, the model using only top five factors almost had the same performance level with the full model, reinforcing the results of machine learning algorithms performed aforementioned.

Table 3
Model loss and accuracy for deep learning sequential model using three optimizers in both training and testing groups.

Optimization algorithms	Training group		Testing group	
	Loss	Accuracy	Loss	Accuracy
All factors				
Adam	9.29%	95.34%	12.38%	93.96%
RMSprop	10.8%	94.78%	12.12%	93.78%
SGD	10.89%	94.67%	11.32%	94.17%
Top 5 factors				
Adam	10.63%	94.32%	11.01%	93.17%
RMSprop	10.56%	94.51%	11.59%	93.12%
SGD	11.02%	94.07%	11.71%	93.33%

Abbreviations: Adam, adaptive moment estimation; RMSprop, root mean square prop; SGD, stochastic gradient descent.

Comparison with traditional Logistic regression model

As the algorithms of both machine learning and deep learning are non-transparent, the traditional Logistic regression model was used to test the association of top five factors selected with the risk of being overweight or obese in children. As shown in Table 4, all five factors were consistently and significantly associated with childhood overweight or obesity at a significance level of 1‰.

Table 4
The risk prediction of top five factors for childhood overweight/obesity using the Logistic regression model.

Top five factors	OR (95% CI)	P
Age of children	1.06 (1.05-1.07)	<0.001
Eating speed	1.43 (1.26-1.62)	<0.001
Number of relatives with obesity	1.25 (1.17-1.34)	<0.001
Sweet drinking	1.15 (1.08-1.22)	<0.001
Paternal education	1.15 (1.06-1.25)	0.001

Abbreviations: OR, odds ratio; 95% CI, 95% confidence interval.

Discussion

The aim of this large survey was to explore the risk profiles of overweight or obesity via a panel of machine learning and deep learning algorithms by analyzing data from children 3-6 years of age from 30 kindergartens. The key finding is the identification of five factors relevant to both children and parents that can better differentiate children with overweight or obesity from the general children, with the prediction performance comparable with that of all factors under consideration. Moreover, it is worth noting that the GBM algorithm is ranked as the best model for the prediction of childhood overweight or obesity. To the best of our knowledge, this is thus far the first study that has employed both machine learning and deep learning techniques to identify and characterize factors associated with childhood overweight or obesity.

Childhood obesity has a complex, multifactorial etiology [25], involving the interactions between inherited and non-inherited factors. Over the last two decades, a vast amount of resources and endeavors have been devoted to identify and characterize factors responsible for overweight or obesity in children [26–32], yet unfortunately the underlying risk profiles still remain to be determined, mainly because the majority of factors identified by traditional statistical analyses have been plagued by inconsistency and non-reproducibility. With the availability of big data, it can be a challengeable task to manually select a traditional regression model (such as linear or logistic regression model) that incorporates perhaps nonlinear and interactive relationships of multiple obesity-susceptibility factors with the health outcome. Moreover, due to the high degree of collinearity among relevant factors, the probability of unstable estimates is high. To help resolve these issues, machine learning and deep learning algorithms can be employed to decipher the complex structure of childhood obesity-susceptibility factors due to excellent predictive power and the capability to handle high-dimensional data.

On the basis of survey data from 9478 children and their parents, we comprehensively appraised the predictive capability of 31 factors for overweigh or obesity among preschool-aged children by means of nine well-known machine learning algorithms, as well as the ensemble classifiers of these algorithms. By gauging five indexes to assess model performance, the GBM algorithm ranked as the best choice. The GBM is a forward learning ensemble methodology, under the rationale that good predictive results can be obtained through increasingly refined approximations, and it builds regression trees on all factors assuming that each tree is built in parallel. Like ours, some studies have shown the excellent predictive performance of the GBM as compared with other models [33, 34]. Besides the identification of best machine learning model, the key finding of this survey was the selection of a minimal number of factors according to their important contributions, including age of children, eating speed, number of relatives with obesity, sweet drinking, and paternal education. The predictive performance of these five factors was comparable with that of all 31 factors under study, as confirmed by the more advanced deep learning technique. The implication of the five factors identified in susceptibility to childhood overweight or obesity is supported by other studies. For example, age is closely related to childhood obesity, as reported by the China Health and Nutrition Survey (CHNS) and China National Nutritional Surveys (CNNS) [35] that the prevalence of overweight and obesity generally increased with age in childhood. Additionally, faster

eating speed was found to be associated with a higher risk of childhood overweight or obesity [14, 36]. Despite the five factors were individually reported to be associated with childhood overweight or obesity, we, for the first time, teased them out from a panel of 31 factors relating to both children and parents. Nevertheless, we agree that further external validations of our findings in other independent populations are necessary, especially by analyzing big data and using artificial intelligence techniques.

Strengths and limitations

This survey is based on a stratified cluster random sampling strategy, and our findings can be extrapolated to local regions. Also, this survey is strengthened by the simultaneous analysis of various sources of data from both children and parents and the adoption of popular machine/deep learning techniques.

Our study has limitations. First, this survey is cross-sectional in design, which cannot address the possible causality effect. Second, childhood overweight or obesity is complex in nature, and only 31 factors were evaluated in this survey. More factors are needed to yield a more reliable estimate. Third, only children of Chinese origin were enrolled, and the extrapolation of our findings to other groups is limited.

Conclusions

We have identified five factors relating to both children and parents that can help differentiate children with overweight or obesity from the general children, and importantly the predictive performance of the five factors was comparable with that of all factors under study. For practical reasons, more studies in longitudinal design, involving a large sample size and adopting more advanced analytical tool, are warranted in the future to characterize the risk profiling of childhood overweight or obesity.

Declarations

Funding

This work received no financial support.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contribution Statement

Z.Z. planned and designed the study and directed its implementation.

Z.Z. drafted the protocol.

Q.W., M.Y., B.P., M.X., Y.Z. and X.D. obtained statutory and ethics approvals.

Q.W., and M.Y, contributed to data acquisition.

Q.W., and W.N. conducted statistical analyses.

Q.W., M.Y., B.P., and M.X., did the data preparation and quality control.

Q.W., and W.N. wrote the manuscript.

All authors read and approved the final manuscript prior to submission.

References

1. Pandita A, Sharma D, Pandita D, Pawar S, Tariq M, Kaul A, Childhood obesity: prevention is better than cure. Diabetes Metab Syndr Obes **9**(83-89 (2016). https://www.ncbi.nlm.nih.gov/pubmed/27042133
2. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. Lancet **390**(10113): 2627-2642 (2017).
3. World Health Organization .Obesity and Overweight. World Health Organization (WHO) (http://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight.Accessed May 28,2018).
4. Lobstein T, Jackson-Leach R, Moodie ML, Hall KD, Gortmaker SL, Swinburn BA, James WP, Wang Y, McPherson K, Child and adolescent obesity: part of a bigger picture. Lancet **385**(9986): 2510-2520 (2015).
5. Jia P, Xue H, Zhang J, Wang Y, Time Trend and Demographic and Geographic Disparities in Childhood Obesity Prevalence in China-Evidence from Twenty Years of Longitudinal Data. Int J Environ Res Public Health **14**(4): (2017).
6. Smego A, Woo JG, Klein J, Suh C, Bansal D, Bliss S, Daniels SR, Bolling C, Crimmins NA, High Body Mass Index in Infancy May Predict Severe Obesity in Early Childhood. J Pediatr **183**(87-93.e81 (2017).
7. Hassink SG, Early Child Care and Education: A Key Component of Obesity Prevention in Infancy. Pediatrics **140**(6): (2017).
8. Benjamin Neelon SE, Østbye T, Hales D, Vaughn A, Ward DS, Preventing childhood obesity in early care and education settings: lessons from two intervention studies. Child Care Health Dev **42**(3): 351-358 (2016).
9. Yan J, Liu L, Zhu Y, Huang G, Wang PP, The association between breastfeeding and childhood obesity: a meta-analysis. BMC Public Health **14**(1267 (2014). https://www.ncbi.nlm.nih.gov/pubmed/25495402
10. Padez C, Mourao I, Moreira P, Rosado V, Long sleep duration and childhood overweight/obesity and body fat. Am J Hum Biol **21**(3): 371-376 (2009). https://www.ncbi.nlm.nih.gov/pubmed/19189418

11. Hunter KE, Johnson BJ, Askie L, Golley RK, Baur LA, Marschner IC, Taylor RW, Wolfenden L, Wood CT, Mihrshahi S, Hayes AJ, Rissel C, Robledo KP, O'Connor DA, Espinoza D, Staub LP, Chadwick P, Taki S, Barba A, Libesman S, Aberoumand M, Smith WA, Sue-See M, Hesketh KD, Thomson JL, Bryant M, Paul IM, Verbestel V, Stough CO, Wen LM, Larsen JK, O'Reilly SL, Wasser HM, Savage JS, Ong KK, Salvy SJ, Messito MJ, Gross RS, Karssen LT, Rasmussen FE, Campbell K, Linares AM, Overby NC, Palacios C, Joshipura KJ, Gonzalez Acero C, Lakshman R, Thompson AL, Maffeis C, Oken E, Ghaderi A, Campos Rivera M, Perez-Exposito AB, Banna JC, de la Haye K, Goran M, Roed M, Anzman-Frasca S, Taylor BJ, Seidler AL, Transforming Obesity Prevention for CC, Transforming Obesity Prevention for CHILDren (TOPCHILD) Collaboration: protocol for a systematic review with individual participant data meta-analysis of behavioural interventions for the prevention of early childhood obesity. BMJ Open **12**(1): e048166 (2022). https://www.ncbi.nlm.nih.gov/pubmed/35058256
12. Liu N, Li H, Guo Z, Chen X, Cheng P, Wang B, Huang G, Shen M, Lin Q, Wu J, Prevalence and Factors Associated with Overweight or Obesity among 2- to 6-year-old Children in Hunan, China: A cross-sectional study. Public Health Nutr 1-32 (2022). https://www.ncbi.nlm.nih.gov/pubmed/35034674
13. Russell SJ, Hope S, Croker H, Crozier S, Packer J, Inskip H, Viner RM, Modeling the impact of calorie-reduction interventions on population prevalence and inequalities in childhood obesity in the Southampton Women's Survey. Obes Sci Pract **7**(5): 545-554 (2021). https://www.ncbi.nlm.nih.gov/pubmed/34631133
14. Liu S, Zhang J, Ma J, Shang Y, Ma Y, Zhang X, Wang S, Yuan Y, Deng X, Niu W, Zhang Z, Synergistic interaction between bedtime and eating speed in predicting overweight and obesity in Chinese preschool-aged children. Aging (Albany NY) **11**(7): 2127-2137 (2019). https://www.ncbi.nlm.nih.gov/pubmed/30978174
15. Zhou B, Yuan Y, Wang K, Niu W, Zhang Z, Interaction effects of significant risk factors on overweight or obesity among 7222 preschool-aged children from Beijing. Aging (Albany NY) **12**(15): 15462-15477 (2020). https://www.ncbi.nlm.nih.gov/pubmed/32741774
16. LeCroy MN, Kim RS, Stevens J, Hanna DB, Isasi CR, Identifying Key Determinants of Childhood Obesity: A Narrative Review of Machine Learning Studies. Child Obes **17**(3): 153-159 (2021). https://www.ncbi.nlm.nih.gov/pubmed/33661719
17. Doupe P, Faghmous J, Basu S, Machine Learning for Health Services Researchers. Value Health **22**(7): 808-815 (2019). https://www.ncbi.nlm.nih.gov/pubmed/31277828
18. Colmenarejo G, Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. Nutrients **12**(8): (2020). https://www.ncbi.nlm.nih.gov/pubmed/32824342
19. Li SM, Ren MY, Gan J, Zhang SG, Kang MT, Li H, Atchison DA, Rozema J, Grzybowski A, Wang N, Anyang Childhood Eye Study G, Machine Learning to Determine Risk Factors for Myopia Progression in Primary School Children: The Anyang Childhood Eye Study. Ophthalmol Ther (2022). https://www.ncbi.nlm.nih.gov/pubmed/35061239
20. Ramos-Gomez F, Marcus M, Maida CA, Wang Y, Kinsler JJ, Xiong D, Lee SY, Hays RD, Shen J, Crall JJ, Liu H, Using a Machine Learning Algorithm to Predict the Likelihood of Presence of Dental Caries

among Children Aged 2 to 7. Dent J (Basel) **9**(12): (2021). https://www.ncbi.nlm.nih.gov/pubmed/34940038

21. Haque UM, Kabir E, Khanam R, Detection of child depression using machine learning methods. PLoS One **16**(12): e0261131 (2021). https://www.ncbi.nlm.nih.gov/pubmed/34914728
22. The Coordinating Team in the Department of Nutrition for Health and Development of the World Health Organization. WHO Child Growth Standards. World Health Organization ((ISBN 92 4 154693 X). 2006: 260–295.).
23. World Health Organization: Child growth standards-BMI-for-age. https://www.who.int/childgrowth/standards/bmi_for_age/en/. (Accessed 11 Nov 2019).
24. de Onis M, Lobstein T, Defining obesity risk status in the general childhood population: which cut-offs should we use? Int J Pediatr Obes **5**(6): 458-460 (2010).
25. González-Muniesa P, Mártinez-González MA, Hu FB, Després JP, Matsuzawa Y, Loos RJF, Moreno LA, Bray GA, Martinez JA, Obesity. Nat Rev Dis Primers **3**(17034 (2017).
26. DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, Fisher CB, Harness D, Viswanath S, Heymsfield SB, Dungan I, Thomas DM, A review of machine learning in obesity. Obes Rev **19**(5): 668-685 (2018).
27. Chatterjee A, Gerdes MW, Martinez SG, Identification of Risk Factors Associated with Obesity and Overweight-A Machine Learning Overview. Sensors (Basel) **20**(9): (2020).
28. Qasim A, Turcotte M, de Souza RJ, Samaan MC, Champredon D, Dushoff J, Speakman JR, Meyre D, On the origin of obesity: identifying the biological, environmental and cultural drivers of genetic risk among human populations. Obes Rev **19**(2): 121-149 (2018).
29. Butler É M, Derraik JGB, Taylor RW, Cutfield WS, Childhood obesity: how long should we wait to predict weight? J Pediatr Endocrinol Metab **31**(5): 497-501 (2018).
30. Robson JO, Verstraete SG, Shiboski S, Heyman MB, Wojcicki JM, A Risk Score for Childhood Obesity in an Urban Latino Cohort. J Pediatr **172**(29-34.e21 (2016).
31. Butler É M, Derraik JGB, Taylor RW, Cutfield WS, Prediction Models for Early Childhood Obesity: Applicability and Existing Issues. Horm Res Paediatr **90**(6): 358-367 (2018).
32. Cortés-Martín A, Colmenarejo G, Selma MV, Espín JC, Genetic Polymorphisms, Mediterranean Diet and Microbiota-Associated Urolithin Metabotypes can Predict Obesity in Childhood-Adolescence. Sci Rep **10**(1): 7850 (2020).
33. Cha GW, Moon HJ, Kim YC, Comparison of Random Forest and Gradient Boosting Machine Models for Predicting Demolition Waste Based on Small Datasets and Categorical Variables. Int J Environ Res Public Health **18**(16): (2021). https://www.ncbi.nlm.nih.gov/pubmed/34444277
34. Atkinson EJ, Therneau TM, Melton LJ, 3rd, Camp JJ, Achenbach SJ, Amin S, Khosta S, Assessing fracture risk using gradient boosting machine (GBM) models. J Bone Miner Res **27**(6): 1397-1404 (2012). https://www.ncbi.nlm.nih.gov/pubmed/22367889

35. Yu DM, Ju LH, Zhao LY, Fang HY, Yang ZY, Guo HJ, Yu WT, Jia FM, Zhao WH, [Prevalence and characteristics of overweight and obesity in Chinese children aged 0-5 years]. Zhonghua Liu Xing Bing Xue Za Zhi **39**(6): 710-714 (2018).
36. Lin M, Pan L, Tang L, Jiang J, Wang Y, Jin R, Association of eating speed and energy intake of main meals with overweight in Chinese pre-school children. Public Health Nutr **17**(9): 2029-2036 (2014). https://www.ncbi.nlm.nih.gov/pubmed/23953989

Figures

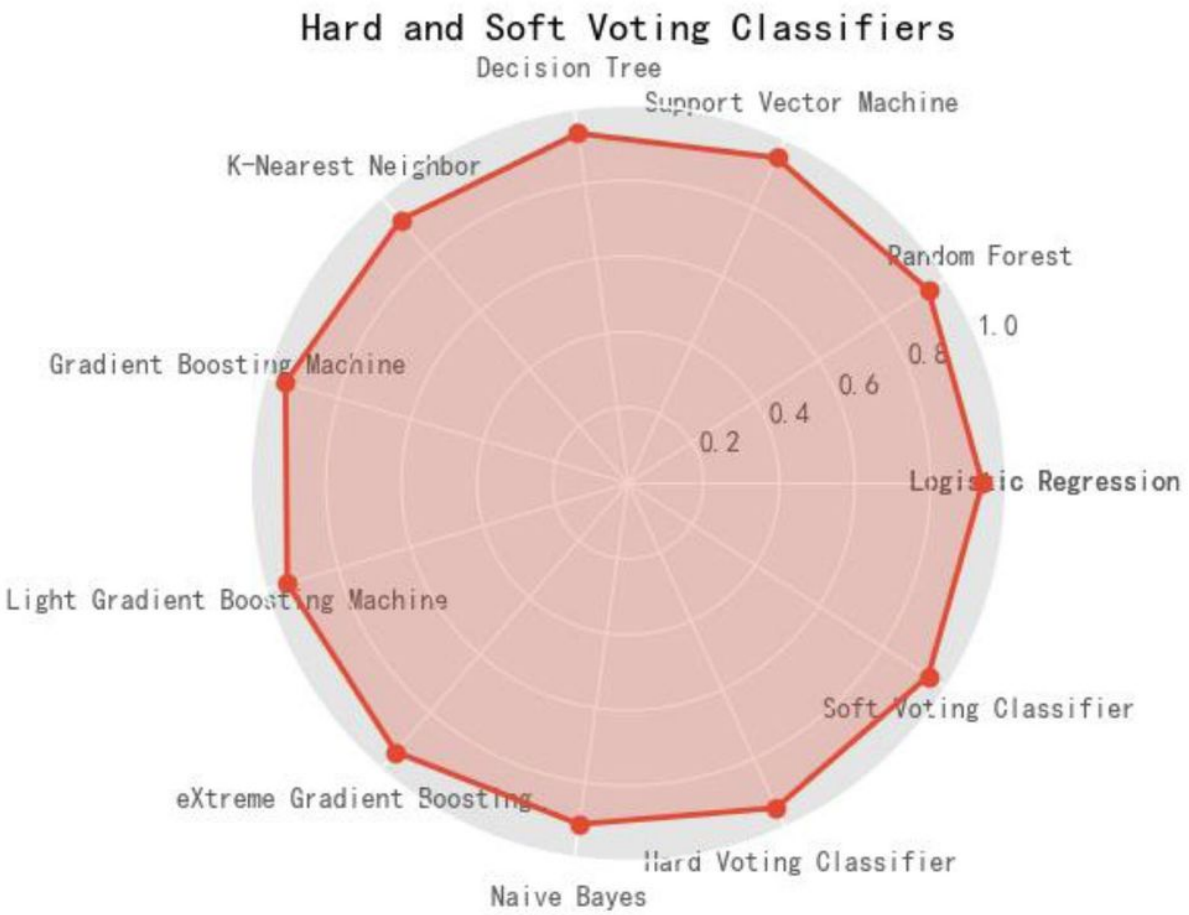

Figure 1

Hard and soft voting classifications based on nine machine learning algorithms for childhood overweight/obesity. Each red solid circle represents the accuracy.

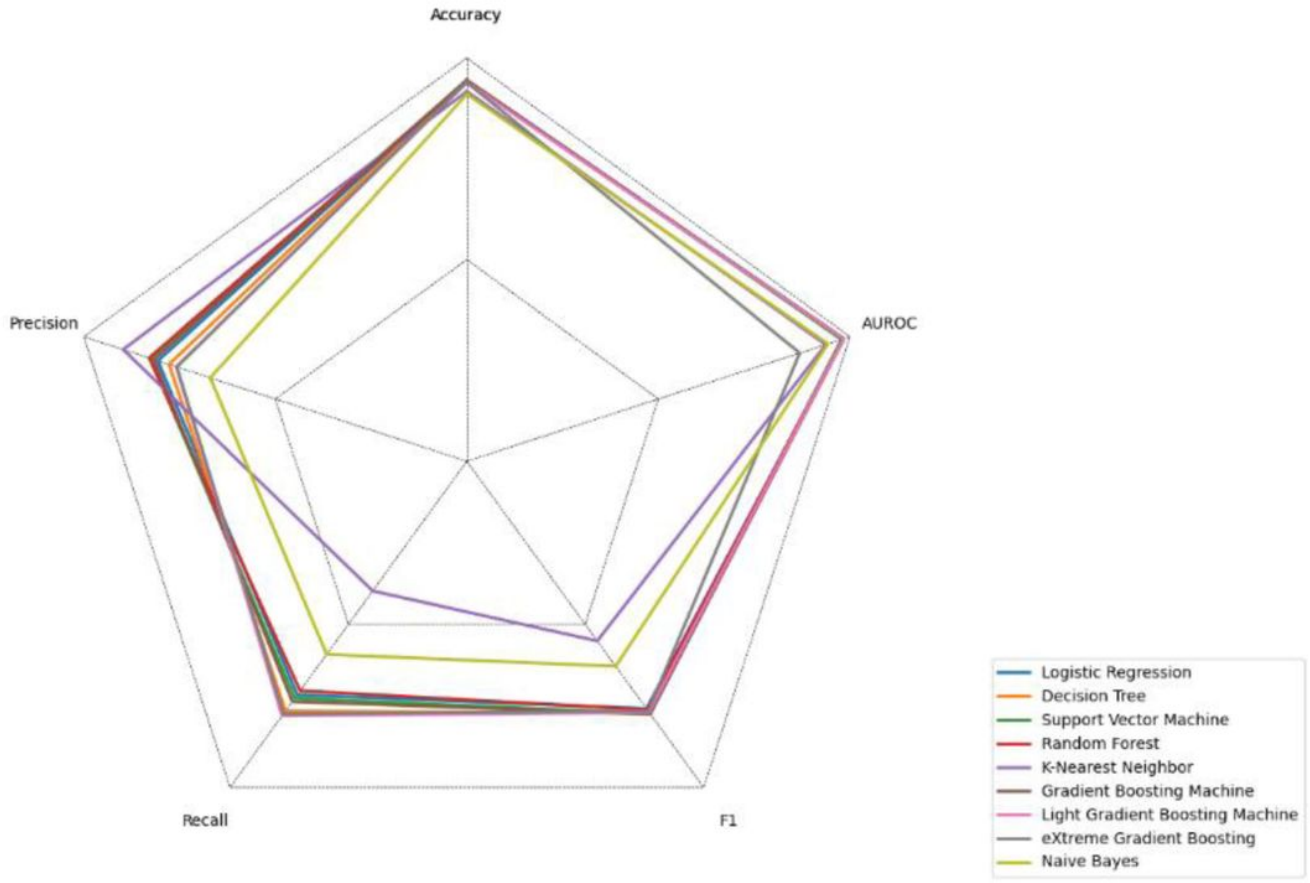

Figure 2

Radar plot illustrating the prediction performance of nine machine learning algorithms in the form of accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (AUROC) for childhood overweight/obesity.

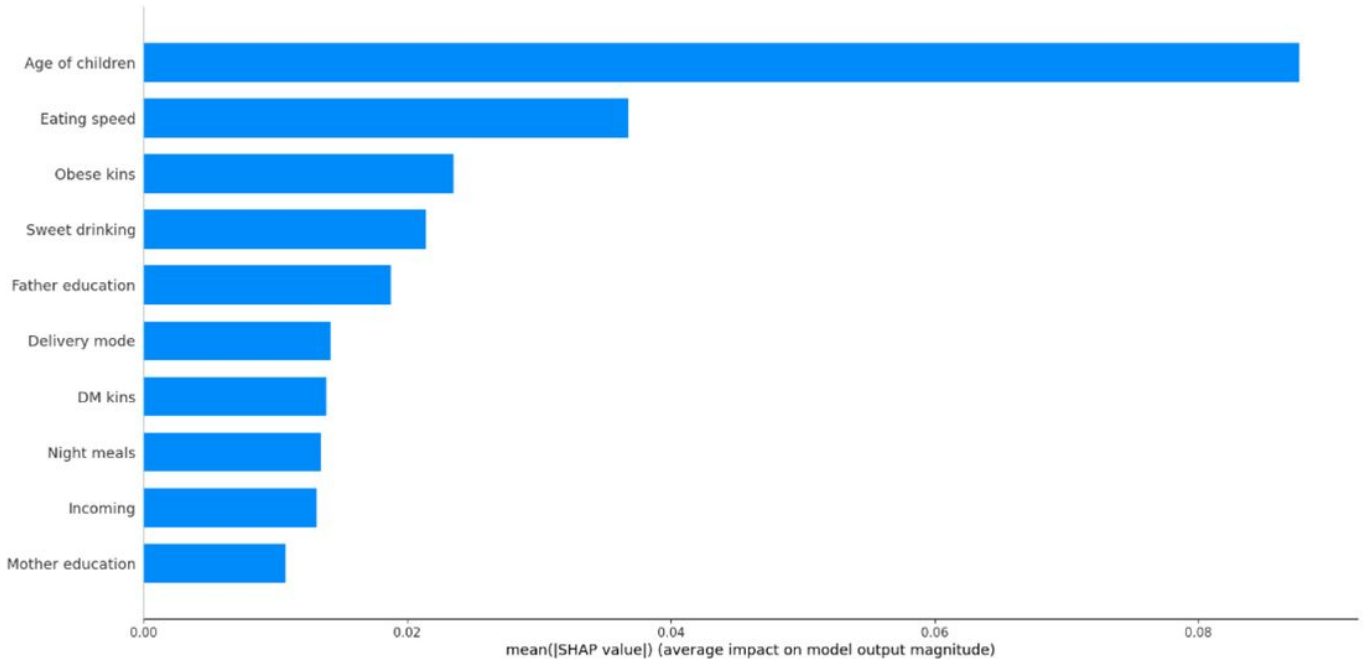

Figure 3

The ranking importance of top ten factors under study for childhood overweight/obesity.