Iran J Public Health, Vol. 54, No.12, Dec 2025, pp.2690-2700

Original Article

Development of a Deep Learning Model for Predicting Obesity Using Health Behavior Data of Elementary School Students

Changgyun Kim [1], ****Ji-Yong Lee*** [2]

1. *Department of Electronic and AI System Engineering, Kangwon National University, Kangwon, Samcheok, 25913, Republic of Korea*
2. *Center for Sports and Performance Analysis, Korea National Sports University, Seoul, Songpa-gu, 05541, Republic of Korea*

***Corresponding Author:** Email: 302479@knsu.ac.kr

(Received 23 Apr 2025; accepted 28 Jun 2025)

Abstract

Background: Childhood obesity poses serious long-term health risks and is a growing global concern. In South Korea, national health surveys collect behavioral and physical data from elementary students, but the large number of questionnaire items can burden young respondents and reduce accuracy. Thus, simplified models with high predictive power are needed.

Methods: We analyzed data from over 250,000 elementary students collected by the Korean Ministry of Education (2015–2022). Using the Rohrer Index as the outcome variable, key predictors were selected via Lasso and Elastic Net regression. Categorical variables were reduced using Multiple Correspondence Analysis (MCA), and a deep learning model (NECTOR) combining MLP and self-attention was developed.

Results: NECTOR achieved high predictive performance with R^2 scores of 0.994 (boys) and 0.996 (girls), and low mean squared errors of 3.072 and 1.841, respectively. It outperformed baseline models using the same inputs.

Conclusion: A small set of core health indicators can effectively predict the Rohrer Index. The proposed model enables efficient and reliable obesity screening in school settings, supporting early intervention efforts.

Keywords: Childhood obesity; Rohrer index; Deep learning; Health behavior; School health survey

Introduction

The prevalence of childhood and adolescent obesity has been increasing rapidly worldwide, not only in high-income countries but also in low- and middle-income nations (1). For instance, the CDC reported that about 20% of U.S. children were obese in 2016. Similar trends are seen in countries like Japan, Australia, and South Korea. Childhood obesity is a major health concern linked to long-term consequences. It significantly raises the risk of adult obesity, which is associated with type 2 diabetes, hypertension, hyperlipidemia, cardiovascular disease, and some cancers (2–3).

The WHO estimates that about 340 million children and adolescents are overweight or obese, with the problem expected to worsen. Obesity is rising even in low-income countries due to globalization, westernized diets, less physical activity,

DOI: https://doi.org/10.18502/ijph.v54i12.20821

and increased screen time (4). Childhood obesity is not just a personal issue but a major public health concern with serious socioeconomic impacts. Consequently, governments and researchers are working on prevention policies, early intervention, and predictive models to address this global challenge.

In this context, the early identification of childhood obesity and at-risk groups has emerged as a critical task for reducing future healthcare costs and preventing disease. In particular, the elementary school years represent a pivotal period during which physical and behavioral habits are formed. Identifying obesity risk and intervening during this stage can significantly prevent the development of chronic diseases in adulthood (5). Therefore, developing an effective and reliable obesity prediction model for elementary school students is of great importance not only from a public health policy perspective but also in clinical and educational contexts.

In South Korea, the Ministry of Education conducts national health surveys to collect data on elementary to high school students' health and behaviors. These surveys cover physical metrics like height and weight, along with various behavioral factors such as diet, physical activity, screen time, and oral hygiene. However, too many questions—often over 25 to 30—can cause fatigue, reduce focus, and lower response reliability (6). Studies show that longer surveys lead to rushed answers and poor data quality, especially among elementary students with shorter attention spans. This increases the risk of misresponses or incomplete data (6).

This study aims to develop a deep learning model that predicts obesity in elementary school students using a minimal set of core variables, addressing the problem of lengthy surveys. Previous studies have shown that childhood obesity can be effectively predicted with fewer variables. For example, one study used LASSO (Least Absolute Shrinkage and Selection Operator) regression on panel data to select 10 key variables and achieved an AUC of 0.82 (7). Another study in Japan used seven binary variables—including sex, screen time, and health conditions—to predict obesity onset, reporting an AUC of 0.803 (8). These findings emphasize that fewer survey items can still yield high predictive accuracy, saving time and resources. Additionally, deep learning models have shown superior performance in this domain. One study using an LSTM model with a self-attention mechanism on EHR data outperformed traditional regression models (9). Deep learning's strength lies in handling complex, high-dimensional data, often surpassing conventional methods in accuracy and scalability (10).

Building on this rationale, the present study aimed to develop a deep learning model capable of predicting obesity in elementary school students using only a small number of core health behavior indicators. By simplifying the complex health behavior questionnaires and reducing the time and cost required for survey completion, the proposed model offers practical value for large-scale obesity screening in school settings—without compromising predictive accuracy. This study employed the Rohrer Index, which accounts for physical development characteristics, as the obesity indicator for elementary school students. The proposed NECTOR (Nominal-Enhanced Correspondence and Transformer-based Obesity Regression) model was constructed by applying dimensionality reduction via Multiple Correspondence Analysis (MCA), followed by a deep neural network using a Multilayer Perceptron (MLP) combined with a self-attention mechanism. The results of this study suggest that the use of the Rohrer Index can help mitigate the limitations of Body Mass Index (BMI) fluctuations during periods of rapid growth and provide a solid foundation for constructing childhood obesity prediction models using only a limited set of input variables.

Methods

Research Data

We used data from the Student Health Examination conducted by Korea's Ministry of Education between 2015 and 2022 (excluding 2020 due to reduced sample size from COVID-19). The da-

taset is a large-scale national panel survey covering physical health and lifestyle behaviors of students from elementary to high school. It includes millions of annual records on height, weight, dietary habits, physical activity, and more.

We focused solely on elementary school students for several reasons. First, this stage marks a rapid acceleration in physical growth, which is strongly linked to future obesity and chronic disease risk (11–12). Second, adolescents show wide individual variation in growth and lifestyle influenced by social and cultural factors, making elementary students a more homogeneous group for analysis (13). Third, the Rohrer Index, which effectively reflects proportional growth in height and weight, is most suitable for this age group (14). Limiting the analysis to elementary students enables clearer insights into how early physical development and behaviors affect obesity. Dataset characteristics are detailed in Table 1.

Table 1: Data Characteristics of Elementary School Students

Year	Male (n)	Female (n)	Total (n)
2015	17,283	16,141	33,424
2016	17,299	16,043	33,342
2017	17,059	16,036	33,095
2018	20,122	18,986	39,108
2019	20,110	18,983	39,093
2020	Excluded from the study		
2021	19,678	18,770	38,448
2022	19,046	18,161	37,207
Total	130,597	123,120	253,717

Feature Selection

In this study, the Rohrer Index, which reflects the relative degree of obesity in growing students, was used as the dependent variable. The Rohrer Index, also known as the corpulence index, was proposed as an alternative measure to complement the limitations of Body Mass Index (BMI) (15). The specific formula is as follows.

$$Rohrer\ Index\ = \frac{Weight\ (kg)}{Height\ (cm)^3} \times 10^7$$

The Rohrer Index—also known as the Ponderal Index or TMI—is widely used to assess childhood obesity, offering a more stable measure than BMI by using height cubed instead of squared (14). Accordingly, this study adopted the Rohrer Index, derived from students' height and weight, as the primary outcome variable.

The dataset included continuous variables (e.g., height, weight) and many categorical items related to diet, physical activity, and home environment. To reduce complexity and avoid overfitting, feature selection was applied. Lasso and Elastic Net regressions, which are effective regularization methods for high-dimensional data, were used to extract key predictors (16–17). In particular, Lasso regression employs an L1 penalty to eliminate irrelevant variables by shrinking their coefficients to zero, enhancing both dimensionality reduction and interpretability (16).

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda ||\beta||_1$$

In the Lasso formula, y_i denotes the dependent variable (locus) for the i-th observation, x_i represents the vector of explanatory variables, and β is the vector of regression coefficients. The term n indicates the number of observations, while λ is the regularization parameter that controls the degree of coefficient shrinkage.

Elastic Net regression is a combination of L1 and L2 regularization, and is particularly effective in handling multicollinearity among variables. This makes it well suited for representing the charac-

teristics of categorical variables. The parameter α adjusts the balance between the L1 and L2 penalties, enabling the model to manage complex relationships among nominal variables in a stable manner. The corresponding optimization formula is presented below.

$$\hat{\beta} = \min_{\beta} \left[\left(\frac{1}{2n}\right) \sum_{i=1}^{n} (y_i - x_i{}^{\mathrm{T}}\beta)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right) \right]$$

This study applied Lasso and Elastic Net regression to identify key explanatory variables affecting the dependent variable. Both are regularization-based techniques that help prevent overfitting and improve interpretability, especially when handling datasets with many categorical and continuous variables. Lasso uses L1 regularization to eliminate irrelevant variables by shrinking their coefficients to zero, while Elastic Net combines L1 and L2 regularization to manage multicollinearity. These models were applied separately for male and female datasets, and the size and direction of coefficients were analyzed to assess each variable's impact. Selected variables are summarized in Table 2.

Table 2: Characteristics of the data

No	Variable name	Description	Scale
1	Ramen	How many times do you usually eat instant noodles in a week?	Categorical (4)
2	Drinks	How many times do you usually drink sweetened beverages in a week?	Categorical (4)
3	Fastfood	How many times do you usually eat fast food (e.g., hamburgers, pizza, fried foods) in a week?	Categorical (4)
4	Meat	How many times do you usually eat meat (e.g., beef, pork, chicken) in a week?	Categorical (4)
5	Dairy	How many times do you usually consume milk or dairy products in a week?	Categorical (4)
6	Fruit	How many times do you usually eat fruit in a week?	Categorical (4)
7	Vegetables	How many times do you usually eat vegetables (excluding kimchi) in a week?	Categorical (4)
8	Breakfast	How do you usually eat breakfast?	Categorical (4)
9	Exercise	Do you exercise more than three times a week to the point of becoming breathless or sweating?	Categorical (2)
10	Sleep	On average, how many hours do you sleep per day?	Categorical (4)
11	Body_image	How do you perceive your body shape compared to your peers?	Categorical (4)
12	Handwashing	I wash my hands with soap before eating or after playing outside.	Categorical (2)
13	Brushing	I brush my teeth more than twice a day.	Categorical (2)
14	Tv_time	I watch television for more than two hours a day.	Categorical (2)
15	Game_time	I use the internet or play games for more than two hours a day.	Categorical (2)
16	Family_support	My family listens to me well and respects my feelings.	Categorical (2)
17	Family_smoking	Someone i live with smokes cigarettes.	Categorical (2)
18	Family_drinking	Someone i live with drinks excessively and it concerns me.	Categorical (2)
19	Lethargy	I feel like everything is bothersome and hopeless.	Categorical (2)
20	Diet_exp	Have you ever tried any of the following to lose weight?	Categorical (4)
21	Height	Measured height	Ratio
22	Weight	Measured weight	Ratio
23	Systolic	Systolic blood pressure	Ratio
24	Diastolic	Diastolic blood pressure	Ratio

Multiple Correspondence Analysis

The selected nominal variables included many subcategories, which, when converted to dummy variables, greatly increased the feature space. With 20 categorical items, one-hot encoding could expand the dataset to hundreds of dimensions, leading to data sparsity and overfitting. To mitigate this, Multiple Correspondence Analysis (MCA) was applied for dimensionality reduction (18). MCA compresses high-dimensional nominal data while preserving relationships among categories, improving model efficiency and interpretability.

$$X = U \sum V^T$$

Here, X refers to the dummy variable matrix of the original data, while U and V represent the row and column coordinate matrices, respectively. These matrices effectively visualize the relationships among the categories of the nominal variables. The explanatory power (inertia) of each dimension is described as follows.

$$Inertia_d = \frac{\lambda_d}{\sum_{i=1}^{p} \lambda_i}$$

Here, λ_d denotes the eigenvalue of the d-th dimension. Dimensions were selected up to the point where the cumulative explained variance exceeded 80%, in order to effectively preserve the essential structure of the data.

To express the relationships more clearly among nominal variables in the reduced space and to quantitatively evaluate their similarity, KMeans clustering was performed. This approach helps reveal structural similarities among nominal variables effectively. The number of clusters was predefined according to the purpose of the analysis. The clustering algorithm can be described as follows.

$$\min_{c} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} ||x_i - c_k||^2$$

In this context, x_i refers to the MCA coordinates of a variable, c_k denotes the center of the k-th cluster, and z_{ik} represents the probability that observation i belongs to cluster k. The clustering results were used to identify groups of similar variables within the MCA space and to guide the subsequent integration of variables. Based on the criterion of visual interpretability, four clusters were applied in this study. To determine the optimal number of dimensions, we measured the Cumulative Inertia, which represents the amount of information explained by the principal components of the variables. This allowed us to assess how many principal components are needed to sufficiently explain the overall dataset. As a result, for the male dataset, the cumulative inertia reached 0.88 with four principal components, indicating a high level of explanatory power. Similarly, in the female dataset, the cumulative inertia was 0.89 with four principal components, also demonstrating strong explanatory ability.5 As the clustered structure explained 88% of the variable variance in both male and female datasets, the original 16 nominal variables were converted into four dimensional values. As a result, the analysis employed a total of eight predictors: four continuous variables and four latent variables derived from the dimensionality reduction of the nominal variables. These were used to train the model to predict the dependent variable.

Construction of a Deep Regression Model Based on a Multilayer Perceptron (MLP) and Attention Mechanism

Based on MCA results, eight variables (four continuous and four nominal) were selected to build a deep regression model for predicting the Rohrer Index. The model employed a Multilayer Perceptron (MLP) to capture nonlinear patterns and incorporated a self-attention mechanism to enhance variable interaction modeling. The architecture included an input layer with eight variables, a dense layer of 64 ReLU-activated units, and a Multi-Head Attention block (two heads, key dimension = 4) to dynamically evaluate variable importance. Designed a hybrid model that combines a multilayer perceptron (MLP) with a self-attention mechanism to improve prediction performance in continuous small-scale data. To

effectively capture the complex interactions between input features, a multi-head attention structure was introduced, and dropout (0.1) and L2 regularization (λ = 0.001) were applied to prevent overfitting. In addition, Layer Normalization and ReduceLROnPlateau callback (factor = 0.5, patience = 10, min_lr = 1e-6) were introduced to improve learning stability and accelerate convergence. The output layer used a linear activation to predict the Rohrer Index. Details of the model's training procedure are shown in Table 3.

Table 3: Model training process

NO	Category	Formula
1	Input Data Embedding	$z^{(0)} = ReLU\left(W^{(0)}x + b^{(0)}\right)$
2	Self-Attention Block	$z^{attn} = LayerNorm\left(z^0 + MultiHeadAttention(z^0, z^0)\right)$
3	Flatten and Forward Propagation through Hidden Layers	$z^{(1)} = ReLU(W^{(1)}Flatten(z^{(attn)} + b^{(1)})$ $z^{(2)} = ReLU(W^{(2)}z^{(1)} + b^{(2)})$
4	Output Layer	$\hat{y} = W^{(3)}z^{(2)} + b^{(3)}$

The Adam optimizer with a learning rate of 0.001 was used for optimization, and the mean squared error (MSE) was adopted as the loss function to maximize the regression performance. Learning was performed for a total of 100 epochs, and the batch size was set to 8. This hyperparameter configuration was set to maintain consistency in experimental conditions for a fair performance comparison with existing comparison models.

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{l} ||W^{(l)}||_2^2$$

Model Evaluation Method

The dataset was split into training (80%) and test (20%) sets to evaluate model performance. Since the results in both training and validation are derived from survey data, imputing missing values could undermine the performance of the analysis; therefore, records containing missing data were excluded. The data were divided separately for males and females (train: 102,870 males, 97,008 females; test: 25,718 males, 24,254 females), and model accuracy in predicting the Rohrer Index was assessed using mean squared error (MSE) and the coefficient of determination (R^2).MSE measures the average squared error between predicted and actual values, while R^2 indicates the proportion of variance explained by the model, with values closer to 1 reflecting higher accuracy. All analyses were conducted using Python version 3.8.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Results

Results of Feature Selection

In the feature selection stage utilizing regularized regression, both Lasso regression and Elastic Net regression were applied independently. As a result, similar key variables were identified for both the male and female student datasets. Table 4 summarizes the regression coefficients estimated using the Lasso and Elastic Net models for the male and female groups. Both methods consistently identified key variables that had a significant impact on the Rohrer Index.

Table 4: Results of Feature Selection

Variable	Male		Female	
	Elastic Net	Lasso	Elastic Net	Lasso
Ramen	-1.010	-1.048	0.004	0.006
Drinks	-0.313	-0.316	-0.012	-0.001
Fastfood	-0.058	-0.053	-0.048	-0.020
Meat	-0.146	-0.153	0.002	0.009
Dairy	-0.127	-0.123	-0.310	-0.302
Fruit	-0.000	-0.002	0.177	0.168
Vegetables	0.397	0.406	0.163	0.152
Breakfast	0.107	0.136	-0.026	0.018
Exercise	0.247	0.240	-0.150	-0.155
Sleep	0.250	0.259	-0.049	-0.963
Body_image	3.941	8.444	2.736	6.254
Handwashing	0.201	0.331	-0.324	-0.328
Brushing	-0.038	-0.079	-0.180	-0.286
Tv_time	-0.050	-0.403	-0.132	-0.586
Game_time	0.038	0.027	-0.048	-0.289
Family_support	0.068	0.071	0.083	0.142
Family_smoking	0.079	0.347	-0.096	-0.390
Family_drinking	0.101	0.138	-0.147	-0.141
Lethargy	0.222	0.231	0.073	0.427
Diet_exp	4.259	5.314	1.888	1.595

For both male and female groups, Lasso and Elastic Net regression identified "body_image" and "diet_exp" as the most influential variables for predicting the Rohrer Index. In the male group, Elastic Net coefficients were 3.941 and 4.259, while Lasso yielded 8.444 and 5.314, indicating a stronger emphasis. Health-related variables like "vegetables," "sleep," and "exercise" also showed consistent positive effects. In contrast, "tv_time," "game_time," and "family_smoking" had minimal impact in Elastic Net but stronger coefficients in Lasso, revealing different interpretations.

For females, similar patterns emerged. "body_image" and "diet_exp" again showed strong effects, while variables like "sleep," "brushing," and "handwashing" had mixed or negative coefficients depending on the model. These findings suggest self-perception and health behaviors are central to body corpulence prediction.

Lasso highlighted key predictors more aggressively but may overlook subtle effects under multicollinearity. Elastic Net provided more stable selection by balancing variable correlations. Based on both models, 20 variables were finalized, including four continuous (height, weight, systolic, diastolic) and 16 nominal variables, excluding "fruit," "game_time," "family_support," and "brushing."

Model Performance Evaluation Results

Table 5 summarizes the performance of four models, including the proposed NECTOR model, which integrates MCA, MLP, and self-attention. All models were trained separately on

male and female datasets (80%) and evaluated using R^2 and MSE. NECTOR outperformed others, achieving $R^2 = 0.994$ (MSE = 3.072) for males and $R^2 = 0.996$ (MSE = 1.841) for females. In contrast, MCA+MLP and MCA+Regression showed lower performance, with MLP+Attention yielding moderate results. These findings confirm that NECTOR effectively captures complex data patterns and offers superior predictive accuracy compared to baseline models.

Table 5: Model Performance Evaluation Results

Variable	NECTOR	MCA+MLP	MCA+Regression	MLP+Attention
Man(R-Squared)	0.994	0.752	0.564	0.667
Man(MSE)	3.072	50.441	101.417	87.291
Woman(R-Squared)	0.996	0.763	0.588	0.691
Woman(MSE)	1.841	48.145	97.526	85.447

Based on the previous results, the NECTOR model was validated using the remaining 20% of the data (male: 25,718 cases; female: 24,254 cases). The regression outcomes are illustrated in Fig. 1 and Fig. 2 for males and females, respectively. While the model generally predicted the Rohrer Index accurately, some deviations were observed for extreme values. These errors were mainly concentrated in groups with abnormally high or low Rohrer Index values, often corresponding to unrealistic height and weight combinations, such as a height of 130 cm with a weight of either 130 kg or 15–20 kg. These anomalies are likely due to input errors or underlying medical conditions. Aside from these outliers, the model achieved high predictive accuracy, with an average error approaching 0.05.

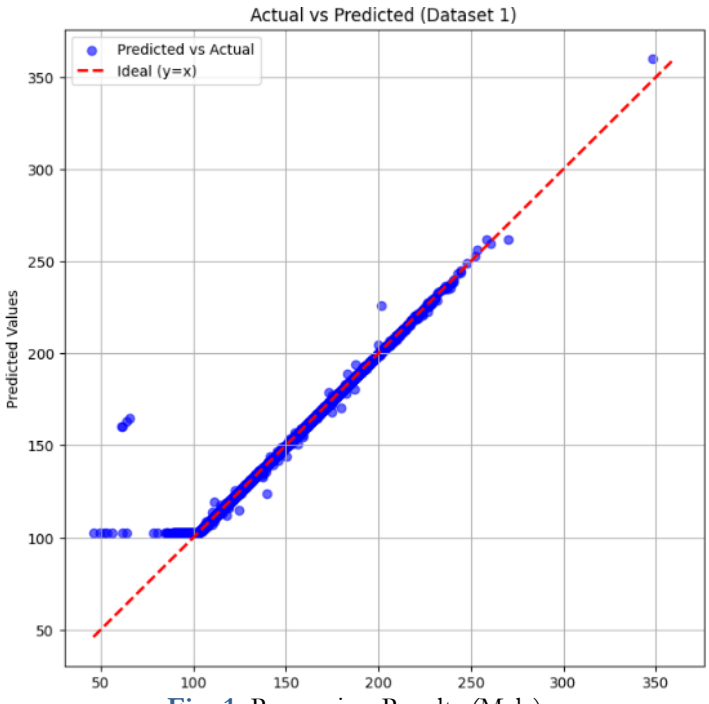

Fig. 1: Regression Results (Male)

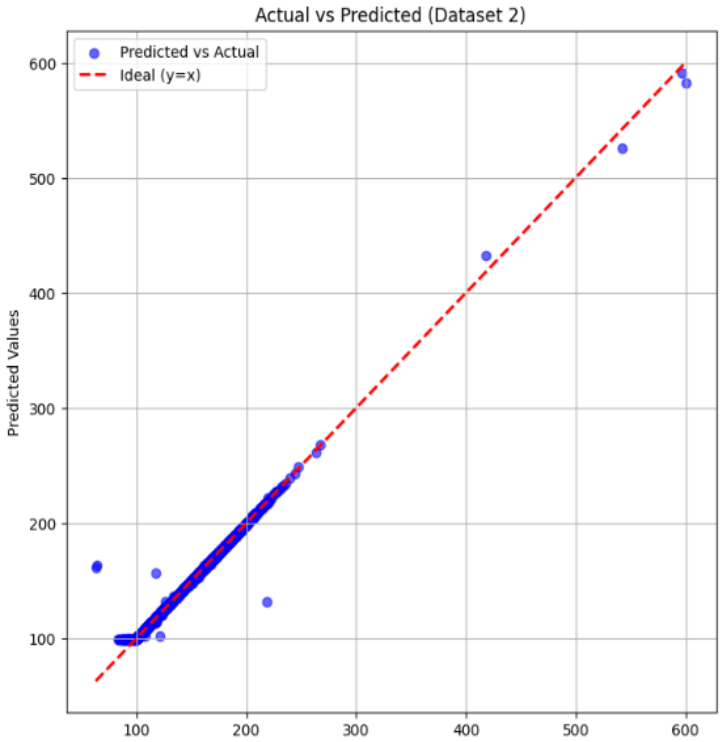

Fig. 2: Regression Results (Female)

Discussion

Beyond predictive performance, the findings of this study highlight important methodological and public health implications for childhood obesity screening. The proposed NECTOR model demonstrates that complex health behavior patterns related to childhood obesity can be effectively captured using a reduced set of core variables, rather than relying on extensive questionnaires. By integrating Multiple Correspondence Analysis with a self-attention–based deep learning architecture, the model preserves latent relationships among categorical health behaviors that are often difficult to represent using conventional statistical approaches. This suggests that dimensionality reduction combined with attention mechanisms can offer an efficient strategy for modeling large-scale health survey data in school-aged populations.

From a public health perspective, this approach is particularly relevant to school health examination settings, where time efficiency, respondent burden, and data quality are critical concerns. The ability to maintain high predictive accuracy while substantially reducing the number of survey items supports the feasibility of implementing rapid obesity risk screening within routine school health programs. Rather than serving solely as a performance-driven modeling exercise, the present findings underscore the potential of data-efficient predictive models to support early identification and prevention strategies for childhood obesity.

The NECTOR model (MCA + MLP + self-attention) outstanding performance in predicting the Rohrer Index, with R^2 scores of 0.994 for males and 0.996 for females. Larger prediction errors were mainly associated with unrealistic height–weight combinations, which are likely attributable to data entry errors or underlying medical anomalies. Excluding such extreme cases, the model demonstrated stable and reliable predictive performance in typical observations.

A major strength of this study lies in its ability to predict childhood obesity with high accuracy using only a small number of core variables. The NECTOR model was trained with a total of 20

variables—four continuous variables and 16 categorical variables—which represents a substantial reduction from the original 112 survey items. Maintaining predictive accuracy despite this reduction offers clear advantages in terms of time efficiency and survey feasibility. Similar findings have been reported by Hammond et al., who developed a model predicting obesity at age five using EHR data and emphasized that minimizing additional data collection can reduce examination costs (19). Consistent with prior research, the present study demonstrates that simplified survey instruments can still yield high predictive performance, thereby reducing respondent fatigue and potential measurement error in school health settings.

However, this study has several limitations that should be considered. First, the analysis was based on data from the Korean Ministry of Education's Student Health Examination, which may limit the generalizability of the findings to other countries or healthcare systems. Validation using external datasets from diverse populations and time periods is therefore warranted. Second, the input variables were restricted to self-reported survey data and basic physical measurements, excluding other potential obesity-related factors such as genetic, environmental, and socioeconomic influences. Third, although the Rohrer Index was adopted as the primary obesity indicator due to its suitability for growing children, alternative metrics such as the Tri-ponderal Mass Index (TMI) have been proposed and should be explored in comparative analyses.

To address these limitations, future research could incorporate longitudinal or sequential modeling approaches, such as recurrent neural networks or graph-based models, to better capture temporal patterns in health behavior data. In addition, integrating broader data domains may enable the development of more comprehensive and generalizable prediction frameworks. From a practical standpoint, the proposed model could support school health professionals by providing rapid obesity risk assessments during routine check-ups, facilitating early referral and tailored intervention. Overall, this study demonstrates that childhood obesity can be predicted efficiently and reliably using a limited set of key variables combined with deep learning techniques, offering a practical foundation for early prevention and health management strategies in pediatric populations.

Conclusion

We developed a deep learning model to predict childhood obesity using elementary students' health behavior data. By selecting key variables through regularized regression and MCA, the model achieved high accuracy with fewer survey items, reducing time and cost. The NECTOR model (MCA + MLP + Self-Attention) predicted the Rohrer Index with R^2 values above 0.99 for both sexes, showing strong potential as a screening tool in school health settings. Given elementary students' limited attention spans, minimizing survey length is crucial. This model offers a practical solution for early obesity detection and may aid in future health management and intervention strategies for children.

Journalism Ethics considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (2024S1A5B5A16024653)

Conflict of interest

The authors declare no conflict of interest.

References

1. WHO (2016). *Ending childhood obesity: Report of the Commission on Ending Childhood Obesity.* https://www.who.int/publications/i/item/9789241510066
2. Bjerregaard LG, Jensen BW, Ängquist L, et al (2018). Change in overweight from childhood to early adulthood and risk of type 2 diabetes. *N Engl J Med,* 378 (14): 1302–1312.
3. Simmonds M, Llewellyn A, Owen CG, Woolacott, N (2016). Predicting adult obesity from childhood obesity: A systematic review and meta-analysis. *Obes Rev*, 17 (2): 95–107.
4. UNICEF (2019). *The State of the World's Children 2019: Children, Food and Nutrition – Growing well in a changing world.* https://www.unicef.org/reports/state-of-worlds-children-2019
5. Herman KM, Craig CL, Gauvin L, Katzmarzyk PT (2009). Tracking of obesity and physical activity from childhood to adulthood: the Physical Activity Longitudinal Study. *Int J of Pediatr Obes,* 4 (4): 281-288.
6. Sharma H (2022). How short or long should be a questionnaire for any research? Researchers dilemma in deciding the appropriate questionnaire length. *Saudi J Anaesth*, 16 (1): 65-68.
7. Lim H, Lee H, Kim J (2023). A prediction model for childhood obesity risk using the machine learning method: a panel study on Korean children. *Sci Rep,* 13 (1): 10122.
8. Sonoda R, Tokiya M, Touri K, Tanomura Y, Yada K, Funakoshi Y, Saito I (2023). A point system to predict the future risk of obesity in 10-year-old children. *Environ Health Prev Med*, 28: 25.
9. Gupta M, Phan TLT, Bunnell HT, Beheshti R (2022). Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. *ACM Trans Comput Healthc,* 3 (3): 32.
10. Colmenarejo G (2020). Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients,* 12 (8): 2466.
11. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH (2000). Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ,* 320 (7244): 1240-3.
12. Dietz WH (1998). Health consequences of obesity in youth: childhood predictors of adult disease. *Pediatrics,* 101(3 Pt 2):518-25.
13. Patton GC, Sawyer SM, Santelli JS, et al (2016). Our future: a Lancet commission on adolescent health and wellbeing. *Lancet,* 387 (10036):2423-2478.
14. Peterson CM, Su H, Thomas DM, et al (2017). Tri-ponderal mass index vs body mass index in estimating body fat during adolescence. *JAMA Pediatr,* 171 (7): 629-636.
15. Khosla T, Lowe CR (1967). Indices of obesity derived from body weight and height. *Br J Prev Soc Med*, 21 (3): 122-8.
16. Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol,* 58 (1): 267-288.
17. Zou H, Hastie T (2005). Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol,* 67 (2): 301-320.
18. Abdi H, Valentin D (2007). Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics,* 2 (4):651-657.
19. Hammond R, Athanasiadou R, Curado S, et al (2019). Predicting childhood obesity using electronic health records and publicly available data. *PLoS One,* 14 (4): e0215571.