Predicting Obesity Prevalence in Bangladesh Using Machine Learning Approach to Demographic and Lifestyle Factors

Nasif Hossain Prio
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
nasif.hp007@gmail.com

Mst.Fahmida Akter
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
fahmidaakter5nov@gmail.com

Md. Mahfujur Rahman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mahfujur@vu.edu.bd

Abdullah Tamim
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tamim.cse.vu@gmail.com

D.M. Asadujjaman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
asadujjaman2207557@stud.kuet.ac.bd

Ahammad Hossain
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
ahammadstatru@gmail.com

***Abstract*— Obesity represents an increasing global health concern, with its incidence escalating swiftly in both industrialized and developing nations. This trend is seen in Bangladesh, where increasing rates among adults and children provide considerable public health challenges. This study analyzes the demographic and lifestyle factors affecting the prevalence of obesity in Rajshahi region, using a dataset of 750 individuals classified into obese and non-obese categories. The study utilized machine learning algorithms, including Binary Logistic Regression (BLR), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM), to identify significant determinants of obesity: age, family history, intake of water, sleep duration, and lifestyle habits. Among these models, RF exhibited the highest prediction accuracy of 96%, underscoring its effectiveness in elucidating complex relationships within the data. The results underscore the necessity of focused treatments and policy reforms to address obesity.**

Keywords: Obesity, Bangladesh, Machine Learning, Public Health, Lifestyle Factors

I. Introduction

Obesity has emerged as a significant public health concern globally in recent years, attributed to the high annual mortality rates. The Centers for Disease Control (CDC) indicated that over one-third of adults are classified as obese[1]. Global adult obesity has more than doubled since 1990 and juvenile obesity has increased fourfold. In 2016, over 18% of children and adolescents aged 5 to 19 were classified as overweight or obese [2]. In 2022, 43% of adults aged 18 and older were classified as overweight, while 16% were categorized as obese [3]. By 2025, the global prevalence of obesity is anticipated to attain 18% in men and 21% in women [4].

In Bangladesh, an increasing trend in obesity is noticed among children, adolescence and adults with a remarkably higher prevalence amidst females compared to males[5]. According to studies, sociodemographic factors and lifestyle behaviors contribute to obesity in urban areas, whereas hypertension is more dominant in rural areas[6]. Obesity is a medical disorder characterized by abnormal or excessive fat accumulation in body, which adversely affects an individual's health[7].

This illness diminishes an individual's quality of life and elevates the risk of chronic diseases, such as diabetes, cardiovascular disease, and specific malignancies. Given the escalating global incidence of obesity, it is essential to understand and address the underlying causes and contributing factors to alleviate its detrimental health effects.

II. Problem Statement

In recent decades, overweight and obesity have emerged as alarming public health issues and pose considerable challenges to the overall healthcare system worldwide[8]. The prevalence of obesity in Bangladesh has increased in both men and women. According to the 2011 Bangladesh Demography and Health Survey (BDHS), 4.6% of adults were obese. The 2017–2018 BDHS found that the prevalence of overweight and obesity was higher than underweight[8]. In 1980, 7% of adults and 3% of children in Bangladesh were overweight or obese. In 2013, those rates had climbed to 17% for adults but only 4.5% for children — according to The Institute for Health Metrics and Evaluation (IHME) of the University of Washington[10].

This study evaluates the efficacy of machine learning algorithms, including the chi-square test, and machine learning models like BLR, DT, SVM, and RF in analyzing obesity-related factors and trends.

III. Literature Review

Overweight and obesity have become critical global health concerns, with rates rising sharply in Bangladesh. The 2017–2018 BDHS reported higher prevalence rates of overweight and obesity compared to underweight, reflecting a significant increase since 2011 [8], [9]. According to studies, Heart Rate Variability (HRV) analysis emerged as a physiological marker for obesity-related patterns linking autonomic nervous system activity to eating behaviors. By combining HRV data with predictive models, researchers have enhanced obesity detection [10].

In recent studies, Gradient Boosting Algorithms have demonstrated higher efficiency in terms of classifying obesity. In particular, CatBoost, XGBoost, and LightGBM have shown enhanced predictive accuracy managing larger datasets without substantial preprocessing[11]. Furthermore, the proposed SVR-EANN hybrid model showed superior performance in predicting Body Fat Percentage (BFP) and identifying significant factors, outperforming benchmark models[12].

These multidisciplinary approaches integrate data-driven insights with behavioral and physiological factors, paving the way for effective obesity management strategies.

IV. Data Collection and Preprocessing

A. *Data Collection*

A questionnaire was developed to collect primary data after reviewing relevant literatures. The survey started with demographic questions, followed by health and lifestyle-related queries. The survey also included a summary of the study's context, purpose, confidentiality agreement, and informed consent. The survey took approximately 5 minutes to complete. The dataset used in this study consists of data collected from individuals from different areas of Rajshahi, Bangladesh, focusing on their eating habits and physical condition to estimate obesity levels.

B. *Data Preprocessing*

Before analysis, the dataset was preprocessed to ensure data quality and compatibility by following the outlined steps. Encoding to ensure formability in the dataset by scaling the data values to a standard range (e.g., 0 for female, 1 for male), handling missing values to address incomplete data entries. Missing data can reduce the effectiveness of the analysis. Ensuring data integrity to maintain consistency and reliability in the dataset. Data cleaning to refine the dataset by removing irrelevant or redundant variables that do not contribute to the study's objectives. These preprocessing steps were implemented using Python libraries such as pandas for data manipulation, and scikit-learn for encoding and scaling. The dataset contains 750 primary records with 16 variables and categorizes the data into 4 classes: Underweight, Normal, Overweight, and Obese (Table I). Furthermore, by combining the classes Underweight, Normal, and Overweight, the Non-obese class is defined (Table II).

C. *Correlation Matrix*

The Correlation Matrix presents the correlations between various lifestyle and demographic variables and only with obese BMI, highlighting significant relationships.

TABLE I. Dataset Detail

BMI Category	Count	Percentage (%)
Underweight	129	17.2
Normal	425	56.67
Overweight	150	20
Obese	46	6.13

TABLE II. Final Classes

BMI Category	Count	Percentage (%)
Obese	46	6.13
Non-Obese	704	93.87

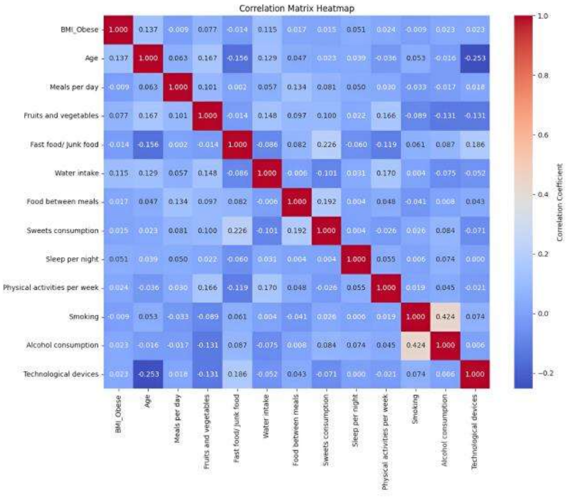

Fig 1. Correlation Matrix

Age is positively correlated with BMI (r=0.137**) and fruit/vegetable intake (r=0.167**), suggesting that older individuals tend to have a higher BMI and consume more fruits and vegetables. Fruit/vegetable intake shows a strong positive correlation with physical activity (r=0.166**) and water intake (r=0.148**), indicating that individuals who consume more fruits and vegetables are also more likely to engage in physical activity and drink more water. Similarly, water intake is positively correlated with both physical activity (r=0.170**) and fruit/vegetable consumption (r=0.148**), further reinforcing the relationship between hydration, diet, and exercise. Physical activity itself is positively correlated with both fruit/vegetable intake (r=0.166**) and water intake (r=0.170**), suggesting that people who are more physically active are also more likely to maintain healthier diet and hydration habits. Finally, technology use is positively correlated with fast food/junk food consumption (0.186**), implying that increased technology usage may be associated with unhealthy eating patterns. These positive correlations highlight the interconnected nature of various lifestyle choices and their influence on health outcomes.

V. Methodology

The Methodology is outlined in a step-wise as follows in the flowchart:

A. *Machine Learning Models*

The study employed machine learning algorithms, including BLR(1), DT(2), SVM(3), and RF(4) to identify significant determinants of obesity, such as age, family history, water intake, sleep duration, and lifestyle habits. Traditional machine learning models were chosen over deep learning methods and neural networks for coherent interpretation of feature importance. These models have demonstrated computational efficiency with structured tabular data. The dataset was split into an 80% training set and a 20% testing set to ensure robust model training and validation for accurate prediction of obesity-related factors.

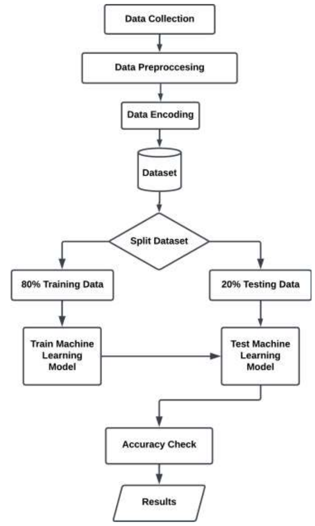

Fig 2. Workflow of methodological steps

$$\boldsymbol{BLR}: \mathbf{P}(\Upsilon\ =\ \mathbf{1}|\mathbf{X}) = \frac{\mathbf{1}}{\mathbf{1}+\mathrm{e}^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \qquad (1)$$

Here,

- P(Υ = 1|X): The probability of the dependent variable Υ being 1 (positive class), given the independent variables $x_1, x_2, x_3 \ldots x_n$.
- β_0: Intercept (bias term).
- $\beta_1, \beta_2, \beta_3$: Coefficients of the independent variables.
- e: The base of the natural algorithm.

$$\boldsymbol{DT}: \boldsymbol{f}(\boldsymbol{x}) = \frac{\mathbf{1}}{|\boldsymbol{R_k}|} \sum_{x_i \in R_k} \boldsymbol{y_i} \qquad (2)$$

Here,

- R_k: The region of the feature space corresponding to the leaf node where x belongs.
- $|R_k|$: The number of data points in the region R_k
- y_i: Target values of the data points in R_k

$$\boldsymbol{SVM}: \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\omega}^{\boldsymbol{T}} \emptyset(\boldsymbol{x}) + \boldsymbol{b} \qquad (3)$$

Here,

- ω^T: Weight vector of the model.
- $\emptyset(x)$: Feature mapping function (possibly non-linear) that transforms x into a higher-dimensional space.
- b: Bias term.

$$\boldsymbol{RF}: \boldsymbol{f}(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{M}} \sum_{\boldsymbol{m}=\mathbf{1}}^{\boldsymbol{M}} \boldsymbol{f_m}(\boldsymbol{x}) \qquad (4)$$

Here,

- M: Number of trees in the random forest.
- $f_m(x)$: Prediction of the m^{th} tree.

B. Evaluation Metrics

Standard accuracy calculation in the classification model is applied to evaluate the performance of the machine learning models: the overall correctness of classification models.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP}+\mathbf{TN}}{\boldsymbol{TP}+\mathbf{TN}+\boldsymbol{FP}+\boldsymbol{F}} \qquad (5)$$

Here,

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

VI. Results and Discussion

Table III presents the analysis highlighting several critical factors influencing obesity, supported by significant Chi-square values and p-values. Age demonstrates a strong association (χ2=39.99, p<0.001), with higher obesity rates observed among older adults, particularly those aged 35 to 60. Family history emerges as a significant predictor (χ2=39.98, p<0.001), with individuals having a genetic predisposition showing markedly higher obesity prevalence. Lifestyle factors such as water intake (χ2=11.54, p=0.002) and sleep duration (χ2=11.4, p=0.003) reveal notable correlations, with participants consuming less than 1 liter of water daily and those with irregular sleep patterns being more prone to obesity. Additionally, consumption of weight-loss supplements is significantly linked (χ2=8.85, p=0.013) to increased obesity rates. These findings underscore the importance of these variables in understanding obesity prevalence.

TABLE III. Characteristics of Participants of Obese

Variable	Category	Non-obese (%)	Obese (%)	χ² and P-Value
Age	<18	15 (100.00%)	0 (0.00%)	χ² = 18.3, p<0.001**
	18-35	641 (94.68%)	36 (5.32%)	
	35-60	41 (80.39%)	10 (19.61%)	
	>60	7 (100.00%)	0 (0.00%)	
Gender	Female	293 (93.31%)	21 (6.69%)	χ² = 0.14, p = 0.702
	Male	411 (94.27%)	25 (5.73%)	
Height	Short	156 (91.23%)	15 (8.77%)	χ² = 6.64, p = 0.036*
	Average	478 (93.91%)	31 (6.09%)	
	Tall	70 (100.00%)	0 (0.00%)	
Family history	No	556 (97.03%)	17 (2.97%)	χ² = 39.98, p =<0.001
	Yes	148 (83.62%)	29 (16.38%)	
Meals per day	1-2 meals	214 (93.45%)	15 (6.55%)	χ² = 0.101, p = 0.951
	Three meals	381 (94.07%)	24 (5.93%)	
	More than three	109 (93.97%)	7 (6.03%)	
Fruits and vegetables	No	64 (95.52%)	3 (4.48%)	χ² = 5.87, p = 0.053*
	Sometimes	295 (96.09%)	12 (3.91%)	
	Yes	345 (91.76%)	31 (8.24%)	
Fast food/ Junk food	No	301 (93.48%)	21 (6.52%)	χ² = 0.05, p = 0.817
	Yes	403 (94.16%)	25 (5.84%)	
Water intake	<1 ltr	64 (96.97%)	2 (3.03%)	χ² = 11.54, p=0.003**
	1-2 ltr	387 (96.03%)	16 (3.97%)	
	>2 ltr	253 (90.04%)	28 (9.96%)	
Food between meals	No	71 (91.03%)	7 (8.97%)	χ² = 1.98, p = 0.370
	Sometimes	445 (94.68%)	25 (5.32%)	
	Frequently	87 (95.60%)	4 (4.40%)	

Sweets consumption	No	249 (93.61%)	17 (6.39%)	$\chi^2 = 1.41$, p = 0.494
	Sometimes	300 (94.94%)	16 (5.06%)	
	Yes	155 (92.26%)	13 (7.74%)	
Sleep per night	<6 hrs.	202 (93.52%)	14 (6.48%)	$\chi^2 = 11.4$, p=0.003**
	6-8 hrs.	447 (95.31%)	22 (4.69%)	
	>8 hrs.	55 (84.62%)	10 (15.38%)	
Physical activities per week	None	361 (93.52%)	25 (6.48%)	$\chi^2 = 1.35$, p = 0.507
	1-2 days	196 (95.61%)	9 (4.39%)	
	3-5 days	67 (95.71%)	3 (4.29%)	
Transportation	Public	517 (94.34%)	31 (5.66%)	$\chi^2 = 2.53$, p = 0.469
	Bike	68 (90.67%)	7 (9.33%)	
	Cycle	14 (100.00%)	0 (0.00%)	
	Walking	55 (93.22%)	4 (6.78%)	
Smoking	No	573 (93.78%)	38 (6.22%)	$\chi^2 = 0.076$, p = 0.963
	Sometimes	29 (93.55%)	2 (6.45%)	
	Yes	102 (94.44%)	6 (5.56%)	
Alcohol consumption	No	660 (94.02%)	42 (5.98%)	$\chi^2 = 0.85$, p = 0.654
	Sometimes	31 (91.18%)	3 (8.82%)	
	Frequently	6 (100.00%)	0 (0.00%)	
Technological devices	0-2 hrs	88 (92.63%)	7 (7.37%)	$\chi^2 = 2.97$, p = 0.226
	3-5 hrs	223 (96.12%)	9 (3.88%)	
	>5 hrs	393 (92.91%)	30 (7.09%)	
Calorie monitoring	No	585 (93.45%)	41 (6.55%)	$\chi^2 = 0.74$, p = 0.388
	Yes	119 (95.97%)	5 (4.03%)	
Supplement consumption	None	663 (94.31%)	40 (5.69%)	$\chi^2 = 8.68$, p=0.013**
	Weight Gain	21 (95.45%)	1 (4.55%)	
	Weight Loss	20 (80.00%)	5 (20.00%)	

Table IV highlights the performance of various machine learning models in predicting obesity. Both BLR and DT models show comparable results, with around 88% accuracy, indicating moderate predictive capabilities. BLR's linear assumptions limit its effectiveness in capturing nonlinear interactions, whereas DT tends to overfit due to reliance on few dominant features. SVM performs better, achieving an accuracy of 92% reflecting its effectiveness in prediction tasks. However, careful feature scaling is required for mixed categorical data. RF stands out as the most accurate model, with a performance of 96%, showcasing its strength in capturing complex relationships in the data, identifying more relevant predictors with a clear assessment of feature importance values compared to others. These results emphasize the effectiveness of advanced ensemble methods, particularly RF, for obesity prediction.

TABLE IV. MODEL PERFORMANCE ON OBESITY PREDICTION

Models	**Accuracy**
BLR	88.67%
DT	88.00%
SVM	92.00%
RF	96.00%

VII. CONCLUSIONS

This study highlights the significant role of demographic and lifestyle factors, such as age, family history, water intake, and sleep patterns, in influencing obesity prevalence in Bangladesh. Using machine learning models, RF achieved the highest accuracy (96%) in obesity prediction, demonstrating its effectiveness. The findings emphasize the need for targeted interventions and broader research to develop in-depth public health strategies to combat obesity in Bangladesh. In the future, we aim to develop a prediction model that provides individuals with comprehensive assessment of their obesity risk by analyzing categorical data and generating a risk measurement score. Therefore, our primary focus will be on collecting high-dimensional datasets from urban areas and employing deep learning models for further robust analysis.

REFERENCES

[1] S. A. G. Ali, H. R. D. AL-Fayyadh, S. H. Mohammed, and S. R. Ahmed, "A descriptive statistical analysis of overweight and obesity using big data," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2022, pp. 1–6.

[2] B. Wu, Y. Jiang, X. Jin, and L. He, "Using three statistical methods to analyze the association between exposure to 9 compounds and obesity in children and adolescents: NHANES 2005-2010," *Environ. Health*, vol. 19, no. 1, 2020.

[3] "Obesity and overweight," *Who.int*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. [Accessed: 13-Jan-2025].

[4] R. An, J. Shen, and Y. Xiao, "Applications of artificial intelligence to obesity research: Scoping review of methodologies," *J. Med. Internet Res.*, vol. 24, no. 12, p. e40589, 2022.

[5] S. M. M. Kamal, C. H. Hassan & G. M. Alam, "Dual burden of underweight and overweight among women in Bangladesh: patterns, prevalence, and sociodemographic correlates." Journal of Health, Population, and Nutrition, 33(1), 92–105, 2015

[6] N. Ali, N. C. Mohanto, S. M. Nurunnabi, T. Haque, and F. Islam, "Prevalence and risk factors of general and abdominal obesity and hypertension in rural and urban residents in Bangladesh: A cross-sectional study," BMC Public Health, vol. 22, no. 1, 2022.

[7] *A Novel Multi-class Classification of Obesity Level using Artificial Neural Network Machine Learning Model*.

[8] M. S. Hossain, N. Tabassum, M. A. Bary, S. J. Shipa, and M. R. Sarkar, "Prevalence of overweight and obesity among Bangladeshi young adults and evaluation of the associated factors after COVID-19 pandemic: A cross-sectional study," *Banglad. Pharm. J.*, vol. 27, no. 2, pp. 182–192, 2024.

[9] T. Kabir, N. N. Popy, and M. S. Alam, "What factors influence consumer overconsumption of food An investigation from Dhaka during the Covid19 Pandemic," *Int. J. Behav. Healthc. Res.*, vol. 8, no. 1, p. 1, 2022.

[10] J. M. Clark, H.-Y. Chang, S. D. Bolen, A. D. Shore, S. M. Goodwin, and J. P. Weiner, "Development of a claims-based risk score to identify obese individuals," *Popul. Health Manag.*, vol. 13, no. 4, pp. 201–207, 2010.

[11] A. Maulana, R. P. F. Afidh, N. B. Maulydia, G. M. Idroes, and S. Rahimah, "Predicting obesity levels with high accuracy: Insights from a CatBoost machine learning model," *Infolitika J. Data Sci.*, vol. 2, no. 1, pp. 17–27, 2024.

[12] Z. Zheng and K. Ruggiero, "Using machine learning to predict obesity in high school students," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 2132–2138.