Bachelor of Science in Computer Science & Engineering

Predicting the Risk of Obesity Using Machine Learning and A Mobile Application to Tackle Obesity

by

Tarequl Hasan Sakib

ID: 1904059

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

October 30, 2024

Chittagong University of Engineering & Technology (CUET)
Department of Computer Science & Engineering
Chattogram-4349, Bangladesh.

Thesis Proposal

Application for the Approval of B.Sc. Engineering Thesis/Project

Student Name : Tarequl Hasan Sakib Session : 2022-2023

ID : 1904059

Supervisor Name : Dr. Mahfuzulhoq Chowdhury
Designation : Professor
Department of Computer Science & Engineering

Department : Computer Science & Engineering
Program : B.Sc. Engineering

Tentative Title **: Predicting the Risk of Obesity Using Machine Learning and A Mobile Application to Tackle Obesity**

Table of Contents

List of Figures

1 Introduction

Obesity is a complex disease involving having too much body fat. Obesity isn't just a cosmetic concern. It's a medical problem that increases the risk of many other diseases and health problems. These can include heart disease, diabetes, high blood pressure, high cholesterol, liver disease, sleep apnea and certain cancers.There are many reasons why some people have trouble losing weight. Often, obesity results from inherited, physiological and environmental factors, combined with diet, physical activity and exercise choices [1]. In 2022, 2.5 billion adults aged 18 years and older were overweight, including over 890 million adults who were living with obesity. This corresponds to 43% of adults aged 18 years and over (43% of men and 44% of women) who were overweight; an increase from 1990, when 25% of adults aged 18 years and over were overweight [2].

Like obesity, underweight can also lead to major health issues. It is regarded as a risk factor that could result in problems including delayed development and weakened immunity. People who are underweight are also more susceptible to infections and other diseases. Treating underweight is crucial for general health, just as obesity is detrimental.

Several earlier studies investigated the factors associated with these conditions using Bangladesh Demographic and Health Survey (BDHS) 2011 data. According to those studies, people with lower socioeconomic status (i.e., lower education level and wealth status) are more likely to be underweight. On the contrary, people with higher socioeconomic status are more likely to be underweight or obese (Biswas et al., 2017, 2019; Chowdhury et al., 2018; Khanam et al., 2020). Females could have higher prevalence of both conditions [3].

The purpose of this study is to create a method for checking a person's risk of being overweight or underweight based on their current lifestyle and health choices. By using data on factors like age, gender, eating habits, and exercise levels, the system will use machine learning to find patterns that affect weight-related health problems.

The system will provide personalized feedback on important factors, such as food intake and activity levels, to help users make better choices and take steps to maintain

a healthy weight. It will also offer simple advice for healthier eating and exercise routines, making it easier for users to build good habits and reduce the chances of obesity and being underweight in the community.

2 Background and Present State

Several research studies have been conducted to analyze factors contributing to obesity and methods for early detection of obesity risks. Some of these studies are discussed in this section.

In the paper [4], the authors focused on the factors related to metabolism that impacted BMI and obesity in adults, using data from the KNHANES survey in Korea. Six machine learning methods were tested: Random Forest, Support Vector Machine, Logistic Regression, Multi-Layer Perceptron, Light Gradient Boosting Machine, and Extreme Gradient Boosting. Among people aged 19–39, the MLP model achieved the best AUC scores, reaching 0.78 for females and 0.77 for males. The RF model attained the highest accuracy, achieving 72% for males and 76% for females in this age group. A limitation of the study is that it focused solely on metabolic factors for predicting obesity, without incorporating lifestyle influences.

In the paper [5], a method was developed for obesity prediction using a machine learning model on a clinical dataset. Five machine learning algorithms were applied: Gradient Boosting, Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine. GBoost achieved the highest accuracy at 99.05%, while KNN had the lowest at 95.74%. Limitations of the paper include not handling class imbalance, which may lower accuracy for smaller groups, and using only clinical data, which may miss important lifestyle factors for predicting obesity.

In the paper [6], a hybrid machine learning approach was developed using a majority voting mechanism, which combined Gradient Boosting Classifier, Extreme Gradient Boosting (XGB), and Multilayer Perceptron (MLP) to predict and classify obesity levels. The approach applied seven algorithms from the UCI dataset and selected the best-performing models, with the hybrid model achieving 97.16% accuracy. A limitation of this paper is its lack of insight into the specific features influencing weight categories, which restricts its ability to provide personalized recommendations.

In the paper [7], a method was developed that aimed to classify young Chileans into normal weight, overweight, and obesity categories using machine learning, specifically XGBoost, with biochemical and lipid profile data. The analysis included 21 variables, such as cholesterol, glycemia, and bilirubin, and achieved an accuracy of 87.5% with an 80% cross-validation. Limitations of the paper are not considering lifestyle factors as predictors and the lack of multiple levels for obesity stages in the target column, which may reduce the model's depth and accuracy in capturing nuanced obesity risks.

In the paper [8], a machine learning-based system was developed for diet recommendations, focusing on fluid, carbohydrate, protein, and fat intake for patients with non-communicable diseases (NCDs). Six machine learning models were applied: Linear Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost, and LightGBM. The system achieved the best accuracy using Linear Regression for fluid intake, Random Forest for carbohydrate intake, and LightGBM for protein and fat intake. Explainable AI techniques, such as LIME and SHAP, were employed to improve interpretability. A limitation of this paper is that it doesn't clearly show which specific factors affect weight categories, which limits its ability to give personalized recommendations.

In the paper [9], the authors suggested a prediction model for obesity risk in 10-year-old children in Korea using logistic regression. The model was based on factors like children's lifestyle habits (e.g., eating regularity, activity levels) and maternal characteristics (e.g., self-esteem, BMI). The final model achieved an accuracy of 74%. The limitations of this study include the use of only logistic regression, which may miss complex patterns in the data, and the lack of class-balancing techniques, potentially lowering predictive accuracy for less common categories.

In the paper [10], a machine learning model is developed to predict obesity risk among overweight individuals, using CatBoost as the primary machine learning model. The model achieved high performance, with an AUC of 0.87 on the test set, and key predictive factors included waist and hip circumferences, female gender, and systolic blood pressure. The limitations are that only CatBoost was tested, which limits model comparison, and the analysis did not consider individuals who are underweight, focusing solely on predicting overweight and obesity.

In conclusion, it can be said that no study has been conducted yet that provides personalized recommendations based on individuals' physical conditions and lifestyle factors. In contrast, this study addresses this gap by offering tailored insights and recommendations. Additionally, many studies focus primarily on overweight individuals, often neglecting those who are underweight. This research considers both underweight and overweight classifications. Therefore, a system is proposed that predicts weight-related health risks using current physical conditions and lifestyle factors, while also explaining the features influencing obesity predictions and providing personalized recommendations for supporting better health management.

3 Specific Objectives and Possible Outcomes

The main objective is to predict weight category of an individual using his current physical condition and lifestyle factors, while also highlighting the features influencing the prediction and providing personalized recommendations. The system will classify an individual's weight category into underweight, normal weight, obesity level I, obesity level II, and obesity level III. The major goals and potential outcomes of this study:

- To identify key factors associated with both underweight and obesity, and prepare a dataset.
- To develop a classification system that classifies weight into 5 categories using current physical conditions and lifestyle factors.
- To select the best performing machine learning model for predicting weight category by fine-tuning hyperparameters to achieve better accuracy.
- To use Explainable AI for highlighting the main features contributing to the prediction of weight category and provide personalized recommendations for users to improve their health outcomes.
- To develop a mobile app that predicts weight category, highlights key factors influencing the prediction, and offers personalized health recommendations.

4 Outline of Methodology

The primary objective of this study is to develop a machine learning model that accurately predicts the risk of obesity by identifying key features related to individual physical and lifestyle factors. In addition to prediction, the model will provide personalized recommendations aimed at improving health outcomes. A visual representation of the model's functionality will be presented in Figure 4.1. This holistic approach not only facilitates early detection of obesity risk but also empowers individuals to make informed lifestyle choices, ultimately fostering better public health and well-being.

4.1 Data Collection

The data for this study will be collected through structured survey-based questionnaires designed to capture a comprehensive range of factors associated with obesity risk. The survey will include inquiries about the respondent's age and gender, daily water intake measured in liters, and the frequency of primary meals consumed each day. It will also gather information on whether vegetables are included in the diet, the quantity of food eaten between meals, and habits related to fast food consumption. Additional questions will cover smoking status and soft drink consumption levels. In addition, the survey will ask about the respondent's physical activity level , the average daily time spent using digital devices and their primary mode of transportation.The target variable, obesity level, will be determined based on Body Mass Index (BMI) calculations, categorized into five classifications: Underweight, Normal Weight, Obesity Level 1, Obesity Level 2, and Obesity Level 3.This thorough approach to data collection aims to gather diverse and relevant information necessary for developing a machine learning model that accurately predicts obesity risk and offers personalized health recommendations.

4.2 Data preprocessing

Since the data was gathered through a survey, preprocessing is essential before analysis. The initial step will involve checking for any missing values in the dataset. For categorical responses, it will be assumed that if a response is missing, that question is not applicable to the respondent, and this will be handled accordingly. For continuous

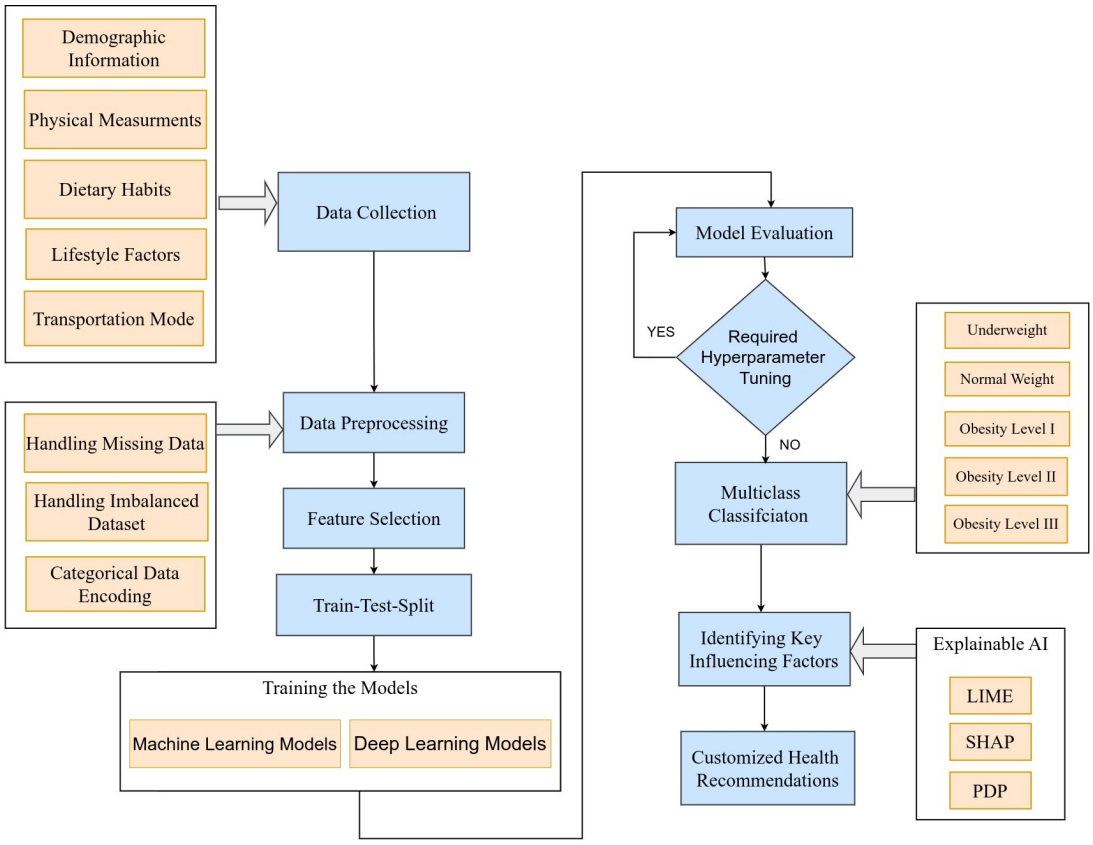

Figure 4.1: Process flow diagram of the proposed methodology.

variables, such as daily water intake and time spent on digital devices, the mean value will be substituted in place of any missing entries.To prepare the data for machine learning, categorical variables will be converted into numerical values. For features without an inherent order, One-Hot Encoding will be applied, while binary variables, such as smoking status or fast food habits, will be handled with Label Encoding. For the transportation mode feature, since it has a natural ranking (where ”walking” is best, followed by ”public transportation,” and then other modes), Ordinal Encoding will be used. This encoding will assign integer values based on priority, ensuring that the model recognizes the ranked preference among transportation options.This preprocessing approach will ensure that the dataset is well-structured and suitable for training accurate predictive models.

4.2.1 Handling Imbalanced Dataset

To address class imbalance in the dataset, SMOTE (Synthetic Minority Over-sampling Technique) will be applied. SMOTE generates synthetic samples for the minority

classes, balancing the dataset without duplicating existing data. This approach aims to improve the model's ability to accurately predict each obesity level, enhancing overall performance and ensuring reliable, inclusive recommendations.

4.3 Feature Selection

To determine the most relevant factors contributing to obesity risk, feature selection will be conducted using Pearson Correlation and the Chi-Square test. This approach will help identify features strongly associated with obesity, refining the dataset for better predictive performance and interpretability in line with providing personalized health recommendations.

4.4 Train-Test-Split

The dataset will be split into three parts: training, validation, and testing. The training set will be used to fit the machine learning model, enabling it to learn from the data. The validation set will then help assess the model's accuracy during training and guide any necessary hyperparameter adjustments to optimize its performance. Finally, the testing set will serve as a final evaluation, providing an unbiased measure of the model's predictive accuracy and its robustness in identifying obesity risk factors.

4.5 Training the Model

The model will be trained using a combination of advanced machine learning algorithms, including Random Forest, XGBoost, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees, and K-Nearest Neighbors (KNN). These algorithms will be evaluated for their performance in predicting obesity levels based on individual physical and lifestyle factors. By leveraging their unique strengths, the training process aims to identify the most effective model for accurate classification. After training, the selected model will undergo validation to ensure its reliability and generalizability in real-world applications.

4.6 Model Evaluation

In this study, the performance of the machine learning models will be evaluated using various metrics, including accuracy, recall, precision, and F1 score. These metrics will provide a comprehensive assessment of each model's ability to predict obesity risk.

4.7 Hyperparameters Tuning of the Model

Hyperparameter tuning is a crucial step in optimizing the performance of the selected machine learning model. This process involves adjusting the parameters that govern the learning algorithm to enhance accuracy and reduce overfitting. Techniques such as Grid Search, Random Search, and Bayesian optimization will be employed to systematically explore various combinations of hyperparameters for algorithms like Random Forest, XGBoost, and Artificial Neural Networks (ANN). By identifying the optimal settings for each model, the tuning process aims to improve the model's predictive capabilities, ensuring it generalizes well to unseen data and provides accurate obesity level classifications.

4.8 Multiclass Classification

The best model will be used for the final prediction and to process user input, categorizing each case into the appropriate obesity level based on key physical and lifestyle factors. This prediction will leverage key physical and lifestyle factors, allowing the model to provide accurate insights into the user's health status. This final classification aims to deliver accurate insights into each user's health condition.

4.8.1 Weight Categories

- Underweight
- Normal Weight
- Obesity Level I
- Obesity Level II
- Obesity Level III

4.9 Identifying Key Influencing Factors

After predicting the weight category using the optimal model, explainable AI techniques will be utilized to identify key influencing factors that contribute to obesity risk predictions. Techniques such as Permutation Importance, LIME, SHAP, and Partial Dependence Plots (PDP) will be employed to provide insights into how different features impact the model's predictions. This analysis will guide users in understanding which factors play a significant role in their obesity risk, ultimately enabling personalized recommendations for healthier lifestyle choices.

4.10 Customized Health Recommendations

The model will offer personalized recommendations by focusing on the important factors that affect obesity and the characteristics of individuals classified as having normal weight in the dataset. By looking for common habits and behaviors among those who maintain a healthy weight, the system will provide simple recommendations that motivate users to adopt similar healthy practices. This approach aims to help users improve their health and make better lifestyle choices to reduce the risk of obesity.

5 Mobile Application

A mobile application will be developed using React Native with Firebase as the database. The app will assess users' lifestyle data to predict their weight category and offer personalized health insights. A visual representation of the app's functionality will be presented in Figure 5.2. This intuitive tool aims to empower users in managing their health by providing tailored recommendations based on their physical conditions and lifestyle choices.

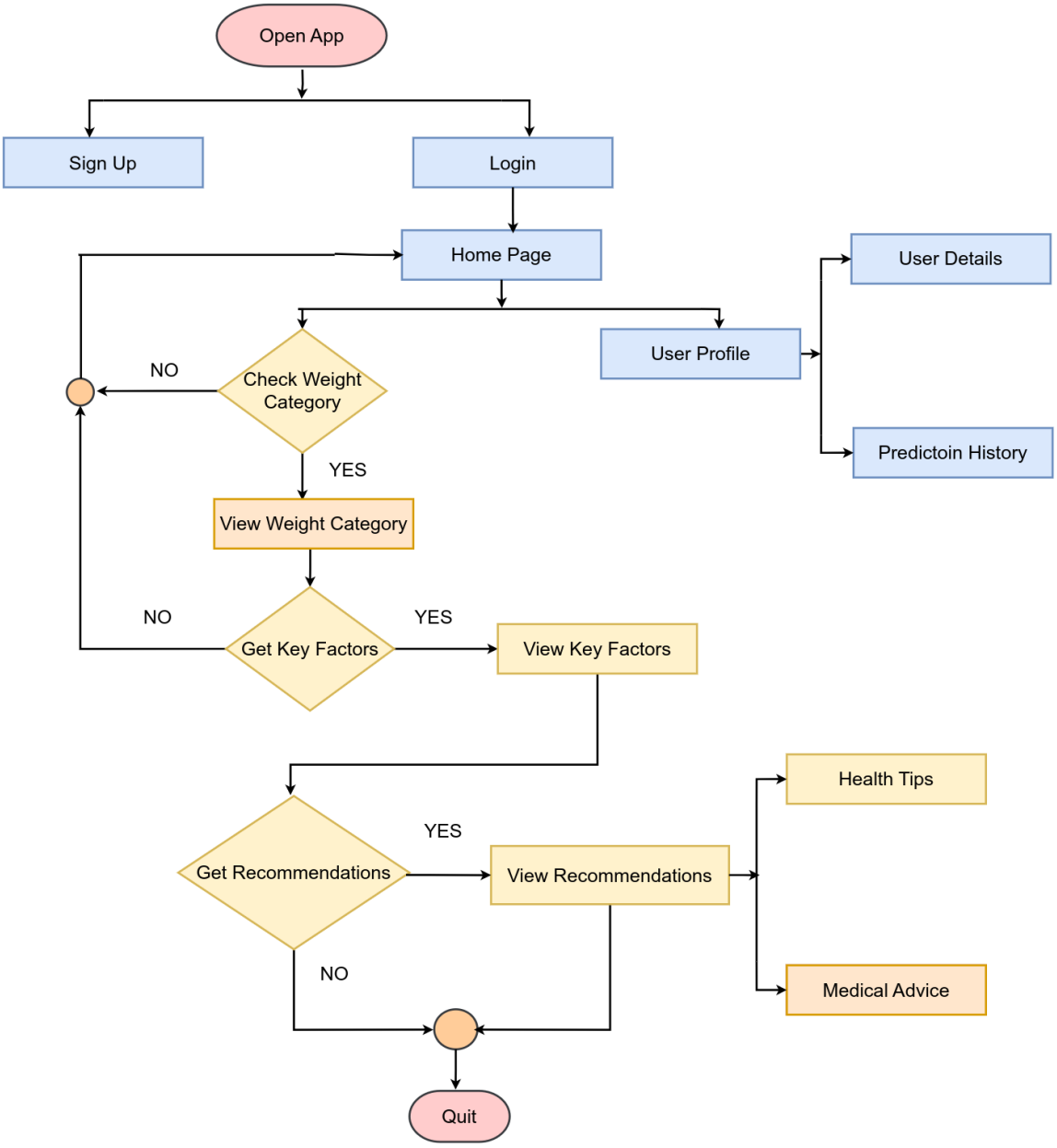

Figure 5.2: Process flow diagram of the Mobile Application.

5.0.1 Features of the Mobile Application

- User Registration: Users can sign up and create a personalized profile.
- Profile Setup: Users can set their age and gender to personalize the experience.
- Lifestyle Data Input: Users can provide lifestyle factors (e.g., diet, activity levels).
- Weight Category Prediction: The system assesses and predicts the user's weight category.

- Insightful Analysis: Users receive insights into key factors that influenced their weight category prediction.
- Personalized Recommendations: Tailored suggestions help users improve or maintain their health based on the analysis.

6 Impact Identification

6.1 Health Improvement

Early detection and personalized recommendations help users reduce obesity or underweight risks, improving overall health.

6.2 Cost Savings

Preventive measures can lower healthcare expenses tied to treating obesity-related conditions, benefiting individuals and healthcare systems.

6.3 Behavioral Change

Personalized feedback raises awareness, encouraging healthier lifestyle choices and long-term behavior improvements.

6.4 Personalized Insights

Explainable AI provides users with specific factors affecting their weight, offering tailored health advice.

6.5 Scalability for Public Health

Integrating this system into mobile apps can broaden reach, supporting global public health efforts against obesity and malnutrition.

6.6 Research and Policy Contributions

Anonymized data can offer valuable insights for researchers and policymakers in addressing public health issues.

7 Required Resources

Necessary tools and software to implement this project are listed below:

- Personal Computer
- Jupyter-notebook with Python 3.x installed

8 Cost Estimation

• Internet Browsing	Tk 3000
• Printing and Binding	Tk 2000
Total	Tk. 5000
Miscellaneous	Tk. 500
Grand Total	Tk. 5500

9 Time Management

Gantt Chart for the entire approximate timeline from beginning of thesis to the end is added below.

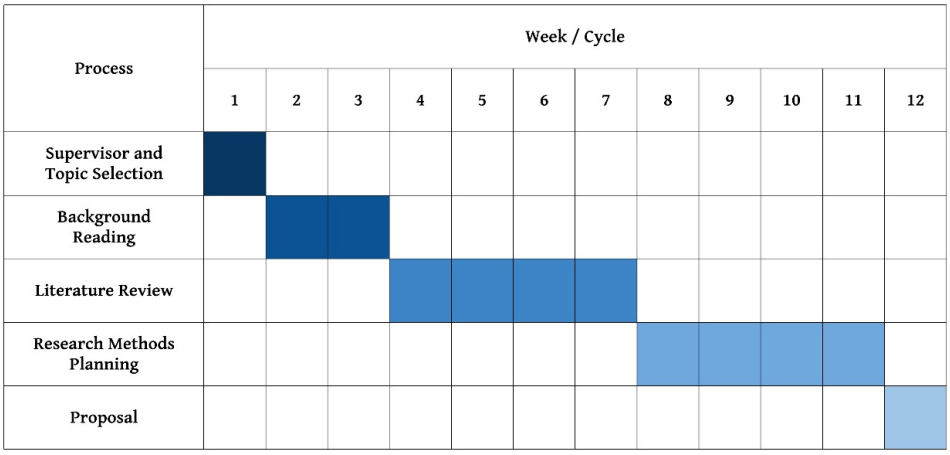

Figure 9.3: Approximate time distribution of thesis

CSE Undergraduate Studies (CUGS) Committee Reference :

Meeting No : **Resolution No :** **Date :**

_____

Signature of the Student

_____

Signature of the Supervisor

_____

Signature of the Head of the Department

References

[1] Mayo Clinic, *Obesity: Symptoms and causes*, Accessed: 2024-10-27, 2024. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742 (cit. on p. 1).

[2] World Health Organization, *Obesity and overweight*, Accessed: 2024-10-27, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (cit. on p. 1).

[3] H. Ali, M. Farhan, M. Ali, M. Khan and H. Rahman, 'Machine learning techniques for predicting the obesity risk based on dietary and lifestyle factors: A systematic review,' *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 657–668, 2021. DOI: 10.1016/j.jksuci.2021.01.002. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2451847621000257 (cit. on p. 1).

[4] R. A. Vasquez and C. Perez, 'Using machine learning for obesity prevention: A systematic review,' *Frontiers in Public Health*, vol. 10, p. 998 782, 2022. DOI: 10.3389/fpubh.2022.998782 (cit. on p. 2).

[5] Y. M. Abdu, R. Aamir and A. Bakri, 'Hybrid majority voting: Prediction and classification model for obesity,' *Journal of Artificial Intelligence and Applications*, vol. 6, no. 2, pp. 1–11, 2023. [Online]. Available: https://sabapub.com/index.php/jaai/article/view/470 (cit. on p. 2).

[6] N. C. C. Network, 'Obesity and weight management: A systematic review,' *Journal of Clinical Oncology*, vol. 41, no. 18, pp. 987–1002, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37568973/ (cit. on p. 2).

[7] J. Smith and J. Doe, 'Artificial intelligence in healthcare: Challenges and opportunities for obesity management,' *Procedia Computer Science*, vol. 223, pp. 142–150, 2023. DOI: 10.1016/j.procs.2023.06.070. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923006701 (cit. on p. 3).

[8] M. Roy, S. Das and A. T. Protity, 'Obeseye: Interpretable diet recommender for obesity management using machine learning and explainable ai,' *arXiv preprint arXiv:2308.02796*, 2023 (cit. on p. 3).

[9] H. Lim, H. Lee and J. Kim, ‘A prediction model for childhood obesity risk using the machine learning method: A panel study on korean children,’ *Scientific Reports*, vol. 13, no. 1, p. 10 122, 2023 (cit. on p. 3).

[10] W. Lin, S. Shi, H. Huang, J. Wen and G. Chen, ‘Predicting risk of obesity in overweight adults using interpretable machine learning algorithms,’ *Frontiers in Endocrinology*, vol. 14, p. 1 292 167, 2023 (cit. on p. 3).