Dr. SUDHA RAJESH
Computational Intelligence
SRM Institute of Science and Technology Chennai, India
sudhar3@srmist.edu.in

PRASHANTH JAVAJI
Computational Intelligence
SRM Institute of Science and Technology Chennai, India
pj3833@srmist.edu.in

Pulaparthi Sathya Sreeya
Computational Intelligence
SRM Institute of Science and Technology Chennai, India
ps0828@srmist.edu.in

Abstract—**Obesity, an excessive body fat condition, has dire health implications. Unhealthy lifestyles, characterized by junk food overconsumption, irregular sleep, and prolonged inactivity, predominantly affect vulnerable adolescents. This intricate medical issue acts as a precursor to various severe ailments like heart disease, strokes, and liver cancer, necessitating immediate attention. As responsible citizens of Bangladesh, addressing this risk is paramount. This research pioneers a machine learning-based approach to predict obesity susceptibility, analyzing a dataset of over 1100 individuals spanning diverse demographics and health conditions. The study evaluates nine prominent machine learning algorithms, including k-nearest neighbor, random forest, logistic regression, multilayer perceptron, support vector machine, naïve Bayes, adaptive boosting, decision tree, and gradient boosting classifier. Through rigorous analysis, the research categorizes obesity risk into high, medium, and low groups, offering valuable insights into individual susceptibility. Combining medical knowledge with technological advancements emphasizes the urgency of combating obesity and deepening our understanding of its intricacies.**

INTRODUCTION

Obesity, the excessive accumulation of body fat, doesn't just stem from dietary factors but also genetics and the environment. Its potential as a global health crisis is concerning for the future. Obesity's roots are complex, leading it to be categorized as a disease due to numerous associated risks. This prevalent health issue is fueled by a sedentary lifestyle and overconsumption of calories. When physical activities can't offset high energy intake, especially from fats and sugars, the surplus energy becomes stored as fat. While some see obesity as a harmless physical trait, it's closely tied to severe illnesses like diabetes, cardiovascular diseases.

The primary focus of this research is to examine individuals for signs of obesity and raise their awareness about the potential risks associated with obesity. The primary objective of this study is to forecast the likelihood of obesity.

The analysis is divided into two primary phases. Initially, the data is ingested and compared against obesity-related factors. Subsequently, the outcomes are presented. To conduct our analysis, we initially gather raw datasets based on specific criteria. These datasets are then preprocessed before being subjected to nine supervised machine learning algorithms. These algorithms are utilized to evaluate metrics such as accuracy, sensitivity, specificity, precision, recall, and F1-score. Through this process, we identify the most effective algorithm for optimal performance in detecting accurate outcomes.

Requirement Gathering

A. User Registration and Authentication:

a. Users should be able to create accounts and log in securely.
b. User roles should be defined (admin, healthcare professional, individual).

B. Data Collection:

a. The system should allow users to input personal and health related data, including age, gender, height, weight, dietary habits, physical activity, sleep patterns, and family history of obesity.
b. Data validation should be implemented to ensure accuracy and completeness.

C. Data Storage:

a. Collected data should be stored securely in a database.
b. Compliance with data protection
c. regulations (e.g., GDPR) is required.

D. Data Preprocessing:

a. Data should undergo preprocessing, including handling missing values, outlier detection, and feature scaling.

E. Machine Learning Models:

a. Implement various machine learning algorithms (e.g., logistic regression, random forest, support vector machine) for obesity risk prediction.
b. Evaluate and select the best performing model(s) based on predefined criteria (e.g., accuracy, F1 Score).

F. Prediction:

a. Allow users to request an obesity risk assessment based on their input data.
b. Display the predicted risk level (e.g., low, medium, high) and associated confidence score.

G. Data Visualization:

a. Provide graphical representations of obesity risk factors and their impact on predictions.
b. Generate charts and reports for users and healthcare professionals.

H. User Feedback and Improvement:

a. Collect user feedback on predictions and user experience.
b. Use feedback to continuously improve prediction accuracy.

I. Security:

a. Implement security measures to protect user data and system integrity.
b. Ensure secure communication (HTTPS) and encryption of sensitive data.

Literature Survey

K. Nirmala Devi et al[1], Machine Learning Based Adult Obesity Prediction.
Techniques:Logistic Regression, Random Forest, and Support Vector Machine, to predict obesity based on factors such as caloric intake, physical activity, genetics, and socioeconomic variables, with Logistic Regression yielding the highest accuracy, titled "Machine Learning Based Adult Obesity Prediction.", K. Nirmala Devi et al. explored A machine learning technique using various algorithms and ensemble methods, like Logistic Regression, Random Forest, and Support Vector Machine, to predict obesity based on factors such as caloric intake, physical activity, genetics, and socioeconomic variables, with Logistic Regression yielding the highest accuracy.

A.S Maria et al[2], Obesity Risk Prediction Using Machine Learning Approach.
Techniques: Gradient Boosting Classifier achieved a 97.08% accuracy in predicting obesity risk using Kaggle's Obesity and Lifestyle dataset, titled "Obesity Risk Prediction Using Machine Learning Approach", A.S Maria et al. explored a machine learning technique employing Gradient Boosting Classifier achieved a 97.08% accuracy in predicting obesity risk using Kaggle's Obesity and Lifestyle dataset.

Sri Astuti Thamrin et al [3], Predicting Obesity in Adults Using Machine Learning.
Techniques: An Analysis of Indonesian Basic Health. In their study published in Volume 8 of 2021, titled "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health," Sri Astuti Thamrin et al. explored the application of machine learning algorithms, specifically Logistic Regression and Naïve Bayesian, on Indonesian health data. Their findings indicated that the Logistic Regression method outperformed Naïve Bayesian, demonstrating superior accuracy, specificity, and precision in predicting obesity in adults based on the Indonesian health dataset. This highlights the potential of machine learning techniques in addressing public health challenges like obesity.

Kamrul H.Foysal et al [4], Predicting Childhood Obesity Based on Single and Multiple WellChild Visit Data Using Machine Learning Classifiers.
In a study conducted by Kamrul H. Foysal, titled "Predicting Childhood Obesity Based on Single and Multiple WellChild Visit Data Using Machine Learning Classifiers," published on January 9, 2023, the researcher explored the prediction of childhood obesity using machine learning techniques. Foysal investigated the performance of six different classification algorithms, including logistic regression, support vector machine (SVM), random forest, artificial neural network (ANN), k-means clustering, and k-nearest neighbors, applied to childhood BMI data. The study assessed multiple machine learning algorithms for each of these scenarios and ultimately recommended models that achieved the highest accuracy, showcasing the potential of machine learning in addressing childhood obesity concerns.

Faray Ferdowsy et al [5], A machine learning approach for obesity risk prediction
In their paper titled "A machine learning approach for obesity risk prediction," authored by Faria Ferdowsy et al. and published in Volume 2, Issue 100053, in November 2021, the researchers emphasized the strong performance of the logistic regression model in the context of predictive work for obesity risk. Their study highlighted the effectiveness and reliability of the logistic regression model in accurately predicting obesity risk, underlining its significance in this machine learning-based approach.

Rajdeep Kaur et al [6], Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence.
In the research conducted by Rajdeep Kaur et al., titled "Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence," published in the Endocrine journal, Volume 78, pages 458–469 in 2022, the authors explored the application of artificial intelligence in predicting the risk of obesity and planning meals to combat obesity in adulthood. They employed machine learning techniques, specifically Gradient Boosting (GB) and Bagging meta estimator (BME), using a dataset sourced from the UCI ML repository. Remarkably, the GB classifier achieved the highest accuracy rate, an impressive 98.11%, underscoring its effectiveness in predicting obesity risk and guiding meal planning strategies to address obesity in adults.

Jocelyn Dunstan et al [7], Predicting nationwide obesity from food sales using machine learning.
In their study titled "Predicting nationwide obesity from food sales using machine learning," conducted by Jocelyn Dunstan et al. and published online on May 19, 2019, the researchers employed three machine learning algorithms for nonlinear regression. These algorithms were applied using data related to food purchase and obesity prevalence. It's important to note that this study utilized machine learning methods, and while it yielded valuable insights, the results should be evaluated with an awareness of both their strengths and limitations, as is common in the application of machine learning techniques in scientific research.

Balbir Singh [8], Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People.

In the study conducted by Balbir Singh, titled "Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People," published in LNTCS, volume 12140, in June 2020, the researcher employed a MultiLayer Perceptron (MLP) as a machine learning approach. They found that the MLP achieved a minority class accuracy of 54% when dealing with an imbalanced dataset. However, this accuracy significantly improved to 92% when the dataset was balanced, demonstrating the importance of addressing class imbalance in predictive modeling.

Kapil Jindal et al [9], Obesity Prediction Using Ensemble Machine Learning Approaches

In the research conducted by Kapil Jindal et al., titled "Obesity Prediction Using Ensemble Machine Learning Approaches," presented at the 5th ICACNI 2018 and published in Volume 2, the authors addressed the challenge of obesity prediction. They highlighted that using only the body mass index (BMI) and obesity equation may not yield accurate results for every individual, as these metrics have limitations. In contrast, their proposed prediction model integrates dual data types, incorporating both regression and classification data, which provides a more comprehensive and accurate approach to obesity prediction.

Erika R. Cheng et al [10], Predicting Childhood Obesity Using Machine Learning: Practical Considerations.

In their paper titled "Predicting Childhood Obesity Using Machine Learning: Practical Considerations," authored by Erika R. Cheng et al. and published in BioMedInformatics in 2022 (Volume 2, Issue 1, pages 184-203), the researchers employed Long Short-Term Memory (LSTM) models to analyze a dataset related to children's Body Mass Index (BMI). Their study focused on the development of machine learning models aimed at identifying pediatric patients who may be at an elevated risk of developing overweight and obesity in the future. This work signifies the practical application of machine learning in addressing childhood obesity concerns and highlights the potential of LSTM models in this predictive context.

Faria Ferdowsy et al [11], A machine learning approach for obesity risk prediction.

In the study conducted by Faria Ferdowsy et al., titled "A machine learning approach for obesity risk prediction," published in Volume 2 in November 2021, the researchers utilized machine learning techniques, specifically logistic regression and the random forest classifier. These algorithms were employed to predict dichotomous measures of obesity risk, categorizing individuals into low obese, medium obese, or high obese groups.

Mathias J Gerl et al [12], Machine learning of human plasma lipidomics for obesity estimation in a large population cohort.

In the study led by Mathias J. Gerl et al., published on October 18, 2019, and titled "Machine learning of human plasma lipidomics for obesity estimation in a large population cohort," the researchers explored the use of machine learning techniques to estimate obesity in a large population cohort based on lipidomic measurements. They found that lipidomic measurements, when combined with machine intelligence modeling, provide a wealth of information regarding both the quantity and distribution of body fat, which goes beyond what can be obtained through traditional clinical assays.

Research Gap/Limitations identified

Data Quality and Availability:

Incomplete Data: Incomplete data is a common issue in healthcare and obesity prediction datasets. Missing values for key variables, such as body mass index (BMI), age, or dietary habits, can reduce the accuracy of predictions. Researchers often use imputation techniques to fill in missing data, but this introduces potential biases.

Inaccurate Self-Reporting: Self-reported measurements, such as self-reported weight, height, or dietary intake, are susceptible to inaccuracies. People may underreport their weight or overreport their physical activity, which can lead to biased predictions. Validation studies comparing self-reported data with clinically measured data are essential to assess accuracy.

Data Collection Methods: The methods used to collect data can impact its quality. For instance, data collected via telephone surveys or online questionnaires may have different levels of accuracy compared to in-person measurements by healthcare professionals. The mode of data collection should be considered when interpreting results.

Generalizability:

Population-Specific Models: Models developed using data from a specific population, such as a particular geographic region or ethnicity, may not be directly applicable to other populations. For example, a model trained on data from one country may not generalize well to a different country due to differences in lifestyle, genetics, and healthcare systems.

Demographic Variability: Demographic factors like age, gender, socioeconomic status, and cultural practices can significantly influence obesity risk and its predictors. Models that do not account for these factors may perform differently across various demographic groups.

External Validation: To assess generalizability, it's important to conduct external validation studies on different datasets or populations. If a model performs well in one population but poorly in another, it indicates a lack of generalizability.

Societal Trends: Societal factors such as changes in diet, physical activity patterns, and sedentary behavior can evolve over time. For instance, shifts toward more sedentary jobs or increased consumption of processed foods can impact obesity rates. Models that do not account for these trends may become less accurate over time.

Cultural Shifts: Cultural norms around body image, food choices, and physical activity can change. For example, a cultural shift toward valuing thinness or adopting a particular diet trend can influence obesity rates. Models should consider the cultural context in which they are applied.

Environmental Factors: Changes in the built environment, such as access to parks, gyms, and healthy food options, can affect obesity rates. Urbanization, for example, can lead to less physical activity and increased reliance on processed foods. Models should incorporate environmental variables when appropriate.

Overfitting: Complex machine learning models can be prone to overfitting the training data, which means they may perform well on the training data but poorly on new, unseen data.

Proposed Methodology

Problem Statement:

Obesity has emerged as a pervasive and pressing public health challenge on a global scale. Its prevalence has reached epidemic proportions, leading to severe health consequences such as diabetes, cardiovascular diseases, and certain types of cancer. Despite the clear need for interventions, identifying individuals at high risk of obesity remains a formidable task. The complexity of this issue arises from the interplay of various factors, including genetics, lifestyle choices, and socio-demographic variables. Without effective means of early identification, preventive measures and interventions often lack precision and fail to target those who need them the most.

Objectives:

The primary objective of our obesity prediction project is to develop a robust predictive model. This model will leverage a comprehensive set of input features, including demographic information, lifestyle data, and genetic markers, to accurately estimate an individual's likelihood of developing obesity. By achieving this, we aim to revolutionize the way we address obesity in healthcare.

1. Early Intervention and Prevention: Our predictive model will enable early identification of individuals at high risk of obesity. This early intervention is pivotal as it permits the implementation of personalized prevention strategies and lifestyle modifications tailored to each individual's unique risk profile.

2. Improving Public Health: By targeting high-risk individuals effectively, we anticipate a significant contribution to the improvement of public health outcomes. A reduction in the prevalence of obesity and associated health issues is a direct outcome we aim to achieve.

3. Resource Allocation Efficiency: Healthcare resources are often stretched thin. Our project will empower healthcare providers by helping them allocate resources more efficiently. Focusing attention on those most likely to develop obesity can alleviate the overall burden on healthcare systems and reduce the economic impact of obesity-related healthcare costs.

4. Awareness and Education: Beyond the predictive model, we aspire to raise awareness about obesity risk factors and promote education on the importance of a healthy lifestyle. Through public education and awareness campaigns, we aim to foster a culture of proactive health management and prevention.

Proposed Architecture Diagram

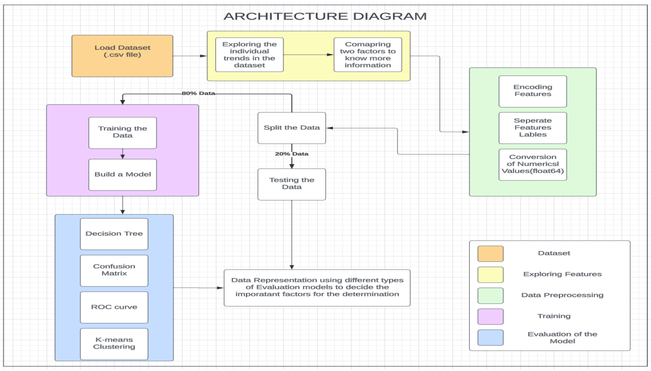


Figure1. Architecture diagram of obesity risk prediction

Figure1 shows the following process:

Exploring features involves understanding the different features in the dataset and their relationships to each other. In Data preprocessing cleaning and preparing the data for training and testing. Encoding features involves converting categorical features into numerical features that can be used by the machine learning model. Training involves feeding the training data to the machine learning model and allowing it to learn the relationships between the different features and the target variable. Evaluation of the model involves using the testing data to evaluate the performance of the trained model.

User: The individual providing their health data and receiving predictions and recommendations. Programmer: Manages system configurations, algorithms, and model updates. System: Represents the entire obesity prediction system. API Layer: Acts as an intermediary between the UI and the backend services, processing requests and communicating with other components. Prediction Engine: The core component responsible for processing the health data and generating obesity predictions using machine learning algorithms.

Modules Description

Exploring features:

Data collection method: The data includes responses from an online survey where respondents had various options to answer each question. Features: Each feature in the dataset holds information collected for specific questions and their corresponding possible answers Categorical variable analysis: To streamline the analysis of categorical variables, a function was created to count and visualize them. This function is designed to prevent repetitive operations and improve efficiency in data exploration. Initial plot: The first plot generated using this function displays the distribution of genders in the dataset. It illustrates the number of men and women as categorical variables.

Design Diagram

Sequence Diagram:

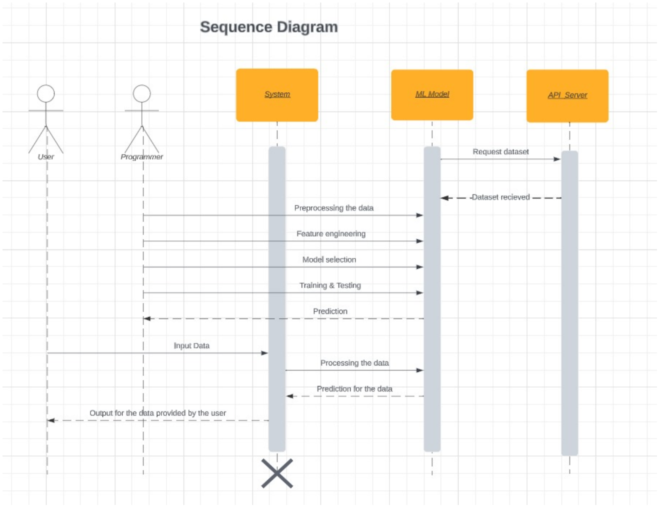


Figure2. Design diagram of obesity risk prediction

Figure2 has the following actors and layers:

Data Preprocessing:

Categorical variable transformation: The categorical variables are transformed using one hot encoding with the `get_dummies()` function. This process creates binary columns for each category within the categorical variables.

Label separation: The labels, which indicate whether a person is overweight/obese or not, are stored separately in a distinct variable. This separation is done to facilitate later use in machine learning algorithms.

Data type adjustment: Upon inspection, it is found that some columns in the dataset contain "float64" numbers while others contain "uint8" values. Machine learning algorithms generally perform optimally with floating point numbers. To ensure uniform data type, all values are converted into float data types.

Feature scaling: It is important that all feature values fall within the same range to prevent issues where the algorithm may misinterpret and assign incorrect coefficients (weights). The features related to obesity are scaled using the MinMaxScaler(), which transforms the values to a range between 0 and 1. Confirmation of the successful scaling is obtained through an examination of the second row of the scaled features.

Label encoding: Most machine learning classification algorithms require labels with numeric values instead of string labels. To address this, the obesity class labels are encoded using LabelEncoder(). The process begins with the instantiation of the encoder, followed by an overview of the data. The `transform()` method is then employed to encode the classes and assign them their respective integer values.

Train and Testing Data:

Data Splitting: The dataset is split into training (80%) and testing (20%) sets. No validation set is retained due to the dataset's small size.

Performance Metric: "F1 score" is chosen as the primary performance metric over "accuracy" for obesity classification, as it considers precision and recall.

Model Selection: DecisionTreeClassifier() is chosen for its simplicity and interpretability.

Hyperparameter Tuning: RandomizedSearchCV() explores tree depths (5 to 15 nodes) to find the optimal combination using 5 fold cross validation, addressing the small dataset issue.

Performance Discrepancy: Training data shows 100% accuracy and F1 score, but testing data drops to around 91%, indicating possible overfitting.

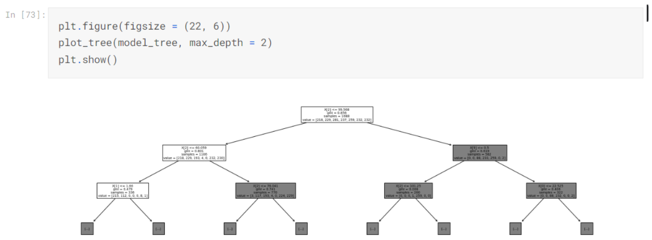
Overfitting Recognized: Overfitting occurs when the model fits training data too closely, making generalization to new data less effective.

Possible Solutions for Overfitting: Options include regularization (e.g., shallower tree, minimum samples per leaf), feature selection, or increasing sample size.

No Further Exploration: Despite overfitting, the model's performance above 90% is considered good, so further adjustments are not pursued at this time.

Results and Discussion

1. Decision Trees for Interpretability:



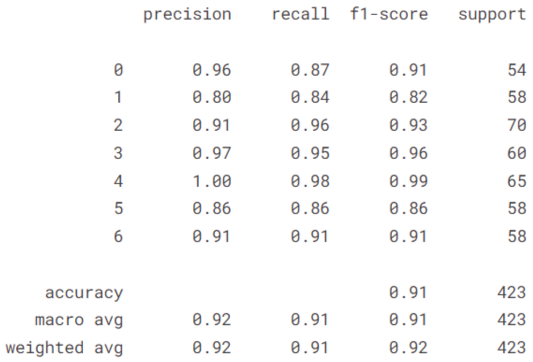
```
In [73]:  plt.figure(figsize = (22, 6))
          plot_tree(model_tree, max_depth = 2)
          plt.show()
```

Figure3. Result of Decision tree

Decision trees are interpretable models.Nodes in the tree ask questions and direct data based on responses (True or False). Max depth limits how many questions can be asked.

2. Classification Report:

Table1. Classification Report (success metrics)



```
              precision    recall  f1-score   support

           0       0.96      0.87      0.91        54
           1       0.80      0.84      0.82        58
           2       0.91      0.96      0.93        70
           3       0.97      0.95      0.96        60
           4       1.00      0.98      0.99        65
           5       0.86      0.86      0.86        58
           6       0.91      0.91      0.91        58

    accuracy                           0.91       423
   macro avg       0.92      0.91      0.91       423
weighted avg       0.92      0.91      0.92       423
```

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.
Recall (Sensitivity): Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to all the actual positives..

Helps understand classification performance.

3. ROC Curve and AUC:

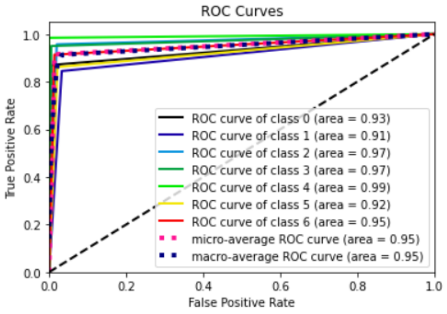


Figure4. ROC Curve for the data

ROC curve plots True Positive Rate vs. False Positive Rate. AUC measures model discrimination ability. AUC ranges from 0 to 1; 0 for random, 1 for perfect predictions. Requires probability prediction scores.

4. k means Clustering Visualization:

Visualizing >3D data on 2D is challenging.Key features (e.g., "Height" and "Weight") used for projection. DecisionTreeClassifier identifies important features. Some columns are not informative and could be removed.

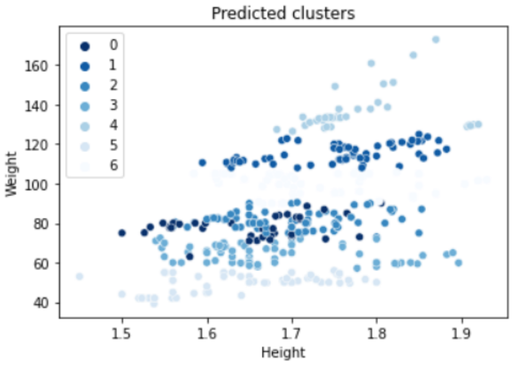


Figure5. Scatter Plot

REFERENCES

1. K. Nirmala Devi, “Machine Learning Based Adult Obesity Prediction”, IEEE Xplore, March 2022.
2. A.S Maria, “Obesity Risk Prediction Using Machine Learning Approach”, IEEE Xplore, May 2023.
3. Sri Astuti Thamrin et al, “Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health”, Frontiers, Volume 8, June 2021.
4. Kamrul H. Foysal et al, “Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers”, National Library of Medicine, Volume 4, January 2023.
5. Faria Ferdowsy et al, “A machine learning approach for obesity risk prediction”, Science Direct, Volume 2, November 2021.
6. Rajdeep Kaur et al, “Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence”, Springer Link, Volume 1, October 2022.
7. Jocelyn Dunstan et al, “Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence”, Sage, Volume 8, May 2019.
8. Balbir Singh et al, “Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People”, Springer Link, Volume 12140, June 2020.
9. Kapil Jindal et al, “Obesity Prediction Using Ensemble Machine Learning Approaches”, Anbar, Volume 2, January 2018.
10. Erika R. Cheng et al, “Predicting Childhood Obesity Using Machine Learning: Practical Considerations”, BioMedInformatics, Volume 2, Issue 1, March 2022.
11. Faria Ferdowsy et al, “A machine learning approach for obesity risk prediction”, Science Direct, Volume 2, November 2021.
12. Mathias J Gerl et al, “Machine learning of human plasma lipidomes for obesity estimation in a large population cohort”, PLOS-Biology, Vol-10,Oc 2019.