ISSN 2277-2685
IJESR/April. 2026/ Vol-16/Issue-2s1/36-42
Mohammed Saifullah *et. al.*, / International Journal of Engineering & Science Research
Full Length Article

Machine Learning–Based Obesity Detection Using Feature-Optimized Xgboost And Comprehensive Evaluation

Mohammed Saifullah[1], Mohammed Abdul Aziz[2], Khaja Asif Uddin[3], Dr. Mohammed Jameel Hashmi[4]
[4]Associate Professor & Head Of Department; Department Of Computer Science Engineering Of ISL Engineering College, Hyderabad, India.
[1,2,3]B.E Student's; Department Of Computer Science Engineering Of ISL Engineering College, Hyderabad, India.
Mail Id: Saifmohammad7742@Gmail.Com, 160522733026@Islec.Edu.In,Aali1503219@Gmail.Com
Accepted 25-04-2026

ABSTRACT:
Obesity has emerged as a major global health concern due to its strong association with chronic diseases such as diabetes, hypertension, cardiovascular disorders, and metabolic complications. Early and accurate prediction of obesity risk is essential for guiding preventive interventions. This project presents a robust machine learning framework for classifying obesity levels using an enhanced XGBoost model trained on a structured lifestyle and physical-condition dataset. The dataset undergoes systematic pre-processing that includes label encoding, normalization, and stratified train-test splitting to ensure reliable learning. The XGBoost classifier is chosen for its superior ability to capture complex feature interactions, handle mixed data types, and reduce over fitting. Comprehensive evaluation metrics such as accuracy, confusion matrix, and classification report demonstrate that the model achieves high predictive performance. Important visualizations, including correlation heatmaps and feature-importance plots, provide deeper insights into the factors influencing obesity outcomes.
The final trained model and processed datasets are saved for deployment and future research. This study highlights the effectiveness of gradient-boosting approaches in health-risk prediction and contributes to the development of intelligent decision-support systems in healthcare.

INTRODUCTION:

Obesity has emerged as a major global health concern due to its strong association with chronic diseases such as diabetes, hypertension, cardiovascular disorders, and metabolic complications. Early and accurate prediction of obesity risk is essential for guiding preventive interventions. This project presents a robust machine learning framework for classifying obesity levels using an enhanced XGBoost model trained on a structured lifestyle and physical-condition dataset. The dataset undergoes systematic pre-processing that includes label encoding, normalization, and stratified train-test splitting to ensure reliable learning. The XGBoost classifier is chosen for its superior ability to capture complex feature interactions, handle mixed data types, and reduce over fitting. Comprehensive evaluation metrics such as accuracy, confusion matrix, and classification report demonstrate that the model achieves high predictive performance. Important visualizations, including correlation heatmaps and feature-importance plots, provide deeper insights into the factors influencing obesity outcomes.

The final trained model and processed datasets are saved for deployment and future research. This study highlights the effectiveness of gradient-boosting approaches in health-risk prediction and contributes to the development of intelligent decision-support systems in healthcare.

LITERATURE REVIEW:

Title: XGBoost vs LightGBM: Gradient Boosting in the Spotlight Author: Data Headhunters Academy
Year: 2024
Description: This study provides an analytical comparison between two leading gradient-boosting algorithms—XGBoost and LightGBM—focusing on performance, speed, memory usage, and efficiency across different machine learning tasks. The article explains how LightGBM achieves faster computation through histogram-based learning and leaf-wise tree growth, while XGBoost offers superior regularization and robustness through exact greedy algorithms. The comparison highlights the strengths and weaknesses of both models, offering insights into their suitability for real-world applications. This survey serves as a valuable foundation for selecting the most effective boosting technique for complex classification problems such as obesity prediction.

Title: Predicting Risk of Obesity and Meal Planning to Reduce Obesity in Adulthood Using Artificial

Intelligence Author: R. Kaur, R. Kumar, and M. Gupta
Year: 2022
Description: This research explores machine learning techniques to predict obesity in adults and generate personalized meal recommendations to reduce obesity risk. The authors use dietary patterns, lifestyle habits, and individual health metrics to develop predictive models such as Gradient Boosting and XGBoost. The study reports high classification accuracy, demonstrating the potential of AI-driven solutions in healthcare and nutrition planning. Through innovative integration of health data and predictive modeling, this work emphasizes how intelligent systems can support early obesity detection and propose practical dietary interventions.

Title: Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living Author: S. Rahman, M. Irfan, M. Raza, K. M. Ghori, S. Yaqoob, and M. Awais
Year: 2020
Description: This paper evaluates the performance of multiple boosting algorithms in recognizing human daily activities using sensor-based datasets. The authors analyze classifiers such as AdaBoost, Gradient Boosting, and XGBoost, comparing their accuracy, stability, and generalization strength. The results show that boosting methods outperform traditional algorithms due to their strong ability to capture complex patterns in activity data. Although not directly focused on obesity, the study demonstrates the effectiveness of boosting models in behavioral and physiological data classification, supporting their applicability in obesity risk prediction systems.

Title: Hybrid Majority Voting: Prediction and Classification Model for Obesity Author: D. D. Solomon et al.
Year: 2023
Description: This study proposes a hybrid majority voting classifier for predicting obesity levels using multiple machine learning algorithms. The model integrates Gradient Boosting, XGBoost, and Multilayer Perceptron (MLP), utilizing a voting mechanism to enhance prediction accuracy. By combining diverse strengths of different classifiers, the hybrid model achieves strong performance on obesity datasets. The research highlights the importance of ensemble strategies in improving classification results and demonstrates that hybrid voting can be an effective solution for multi-class obesity prediction tasks.

Title: Comparison of Prediction of Obesity Status Based on Different Machine Learning Approaches with Different Factor Quantities
Author: G. Shao Year: 2022
Description: This paper compares various machine learning models, including Decision Trees, SVM, XGBoost, and Random Forest, for predicting obesity status using lifestyle and physical condition parameters. The study emphasizes how different numbers of input factors influence model performance and accuracy. Among the tested models, XGBoost demonstrates superior performance due to its stronger feature handling and reduced overfitting. The study provides useful insights into selecting the most suitable algorithm and feature set for obesity prediction, reinforcing the effectiveness of boosting-based approaches.

METHODOLOGY:

The core mathematical principles used in this obesity prediction project are:

1. Data Preprocessing & Linear Algebra
Before the model can "learn," the data must be converted into a mathematical format (vectors and matrices).

- Min-Max Scaling (Normalization): This technique scales numerical features (like Age or Weight) to a fixed range, usually $[0, 1]. This ensures that features with larger magnitudes don't dominate the model.

x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}

- Label Encoding: Categorical data (like "Gender" or "Mode of Transportation") is mapped to integers. This is a simple discrete mapping f: S \rightarrow \mathbb{Z} that allows the algorithm to perform operations on text-based data.

2. Synthetic Data Generation (SMOTENC)
Since datasets are often imbalanced (some obesity levels might have fewer samples), the project uses SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features).

- Geometry & Euclidean Distance: SMOTE identifies the k-nearest neighbors of a minority class point using distance formulas.
- Linear Interpolation: It creates new "synthetic" points along the line segment connecting two existing points in the feature space:

x_{new} = x_i + \lambda \times (x_{neighbor} - x_i)

where λ is a random number between 0 and 1.

3. Gradient Boosted Trees (XGBoost)
The heart of the project is the XGBoost algorithm, which is based on Gradient Boosting.

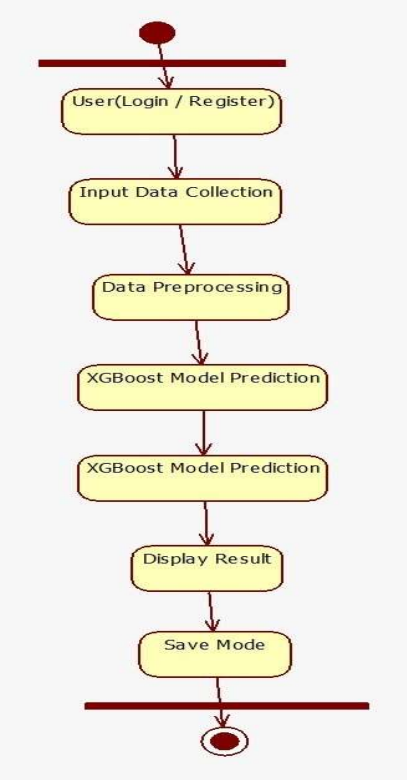
• Calculus (Optimization): XGBoost minimizes a "Loss Function" (how far the prediction is from the truth) using Gradient Descent. It uses a Second-order Taylor Expansion to approximate the loss function, making it much faster and more accurate than standard boosting.
• Objective Function: The model optimizes a combination of a loss function (L) and a regularization term (\Omega) to prevent overfitting:
\text{Obj}(\theta) = \sum L(y_i, \hat{y}_i) + \sum \Omega(f_k)
• Softmax Function: For multi-class classification, the model uses the Softmax function to turn raw scores into probabilities that sum to 1:
P(y=i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}

4. Model Explainability (LIME)

LIME (Local Interpretable Model-agnostic Explanations) is used to explain why a specific prediction was made.
• Local Linear Approximation: LIME assumes that while the global model is complex (non-linear), it can be approximated locally around a single data point using a simple Linear Regression model.
• Kernel Functions: It uses a mathematical "kernel" to give more weight to synthetic samples that are closer to the actual instance being explained, ensuring the explanation is locally faithful.

5. Evaluation Statistics

Finally, the project evaluates performance using several statistical metrics:
• Accuracy: The ratio of correct predictions to total predictions.
• F1-Score: The harmonic mean of Precision and Recall, which is particularly useful for balanced evaluation: F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
• Log Loss: A measure of how close the predicted probability is to the actual value (0 or 1).



By combining these concepts, the project transforms raw lifestyle data into a sophisticated, interpretable diagnostic tool.

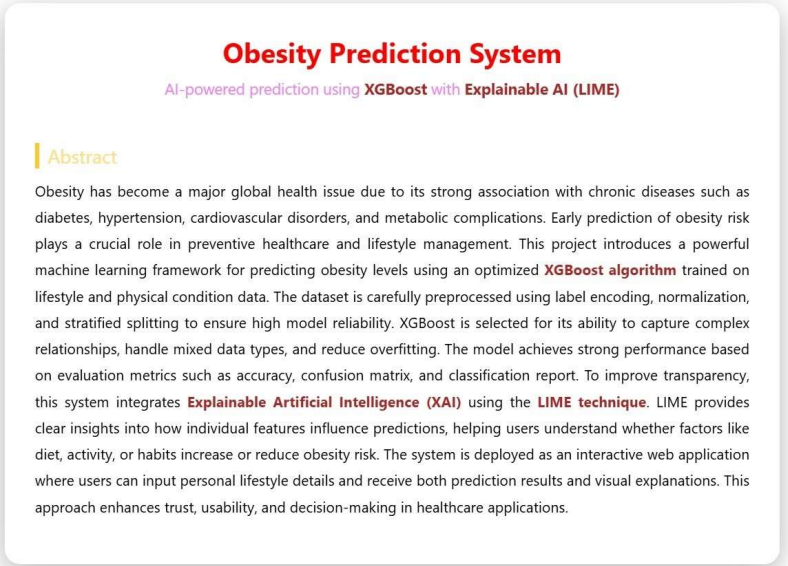
IMPLEMENTATION:

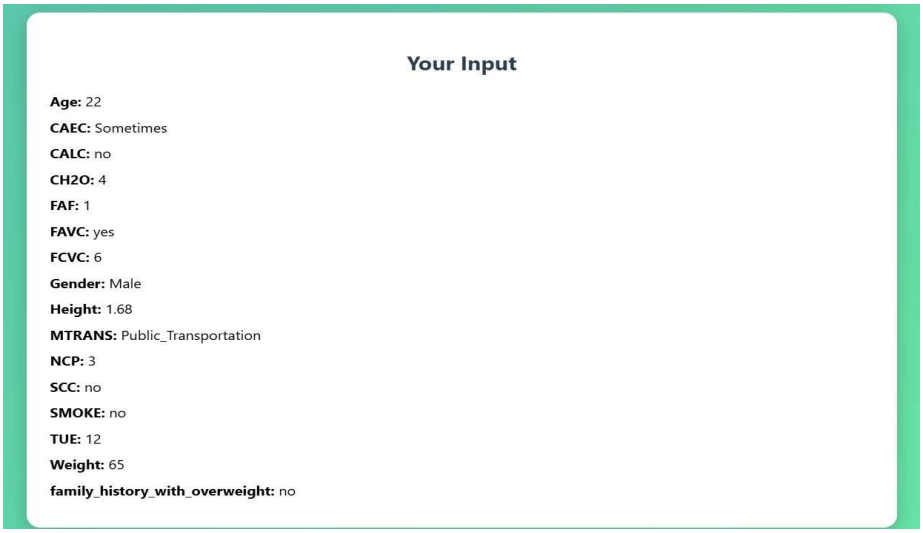
Algorithm 1: Data Preparation and Model Training

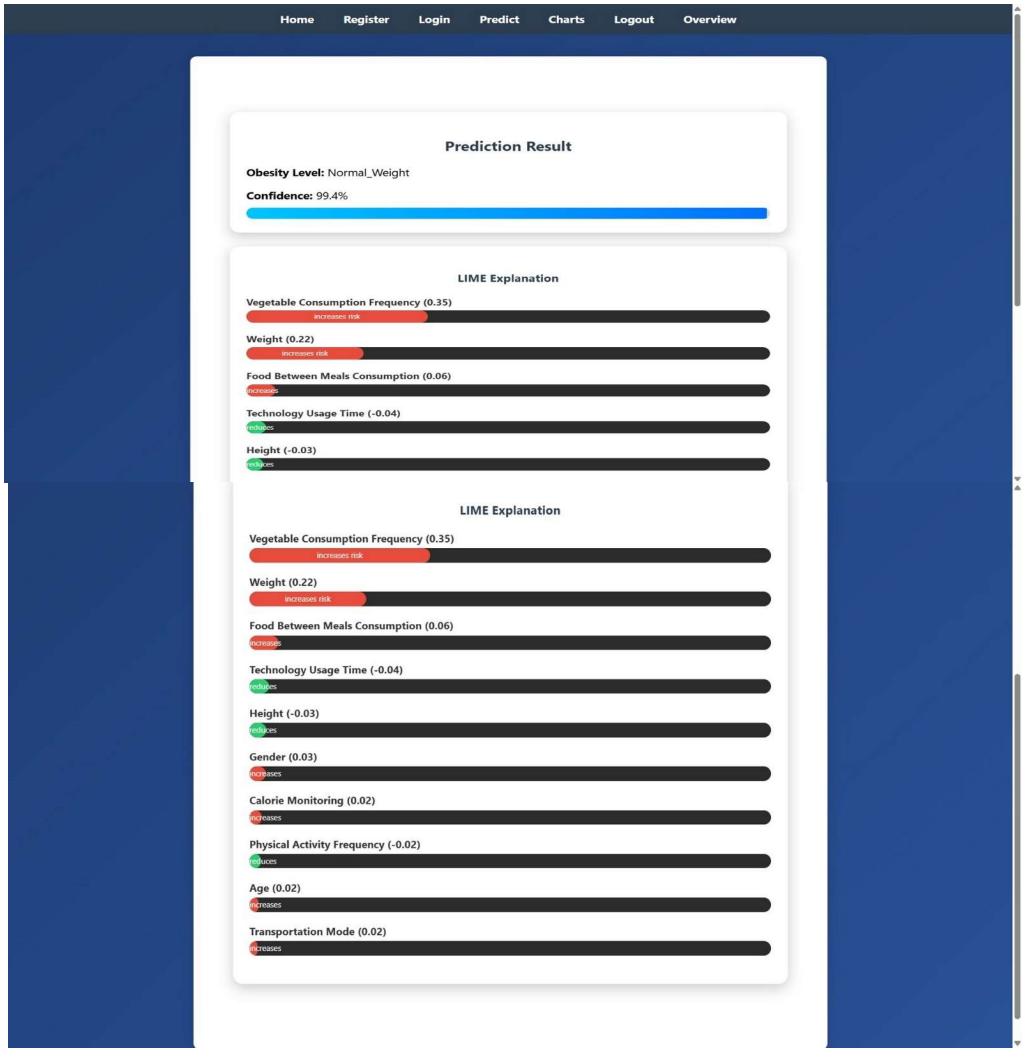
1. Load the obesity dataset from the csv file into a dataframe.
2. Separate the target variable which is the obesity level from the input features.
3. Identify features as either categorical like gender or numerical like age and weight.
4. Apply label encoding to categorical features to transform text labels into unique integers.
5. Apply min-max scaling to numerical

features to bring all values into a range between zero and one.
6. Split the data into a training set for learning and a testing set for evaluation.
7. Use smotenc to generate synthetic samples for minority classes to ensure the dataset is balanced.
8. Define the xgboost parameters including the multi-softprob objective for classification.
9. Train the xgboost model by iteratively building decision trees that correct previous errors.
10. Export the trained model, the feature encoders, and the scaler as files for use in the web app. Algorithm 2: Web Application Prediction and Explanation
1. Start the flask server and load the pre-trained xgboost model and transformation files.
2. Receive user input through the web form for lifestyle habits and physical metrics.
3. Transform the user input using the loaded label encoders for text and the min-max scaler for numbers.
4. Arrange the processed data into a dmatrix format required by the model.
5. Run the model on the input to calculate probability scores for every obesity category.
6. Identify the category with the highest probability as the final prediction.
7. Initialize the lime tabular explainer using the original training data statistics.
8. Generate a local explanation by perturbing the user input and seeing how the model reacts.
9. Extract the specific feature weights that show which habits increased or decreased the risk.
10. Send the prediction result and the visual lime data to the results page for the user to view.

RESULTS:




Your Input
Age: 22
CAEC: Sometimes
CALC: no
CH2O: 4
FAF: 1
FAVC: yes
FCVC: 6
Gender: Male
Height: 1.68
MTRANS: Public_Transportation
NCP: 3
SCC: no
SMOKE: no
TUE: 12
Weight: 65
family_history_with_overweight: no


Home
Register
Login
Predict
Charts
Logout
Overview
Prediction Result
Obesity Level: Normal_Weight
Confidence: 99.4%
LIME Explanation
Vegetable Consumption Frequency (0.35)
increases risk
Weight (0.22)
increases risk
Food Between Meals Consumption (0.06)
Technology Usage Time (-0.04)
Height (-0.03)
LIME Explanation
Vegetable Consumption Frequency (0.35)
increases risk
Weight (0.22)
increases risk
Food Between Meals Consumption (0.06)
Technology Usage Time (-0.04)
Height (-0.03)
Gender (0.03)
Calorie Monitoring (0.02)
Physical Activity Frequency (-0.02)
Age (0.02)
Transportation Mode (0.02)

CONCLUSION:

In conclusion, this project successfully demonstrates the effectiveness of machine learning—particularly the XGBoost algorithm—in accurately predicting obesity risk levels based on lifestyle, demographic, and physical attributes. Through systematic pre-processing, feature encoding, normalization, and rigorous evaluation, the model achieves strong classification performance and provides valuable insights into the factors influencing obesity. The use of visualizations and feature-importance analysis enhances interpretability, making the system practical for healthcare professionals, nutritionists, and individuals seeking early health assessments. The project lays a solid foundation for intelligent health analytics and showcases how data-driven approaches can assist in monitoring and preventing obesity-related conditions. With future enhancements such as deeper models, real-time data integration, and full deployment through web or mobile platforms, this system has significant potential to evolve into a comprehensive and reliable decision-support tool for personalized healthcare.

FUTURE SCOPE:

Future enhancements for this obesity prediction system aim to further improve accuracy, scalability, and real-world usability. One potential improvement is the integration of more advanced deep learning models such as TabNet or Transformer-based architectures to capture deeper relationships within lifestyle and health data. The system can also be expanded to support larger and more diverse datasets, enabling better generalization across different age groups, regions, and ethnicities. Incorporating wearable device data, such as physical activity logs and real-time calorie monitoring, would transform the model into a dynamic health-tracking tool rather than a static predictor. Additionally, implementing advanced explainable AI techniques like SHAP values could provide deeper insights into feature contributions and improve trust among healthcare professionals. Finally, future versions may include a full-scale mobile or web-based platform with user profiles, progress dashboards, and personalized recommendations, making it a comprehensive digital health assistant for obesity management and prevention.

REFERENCES:

[1] R. Kaur, R. Kumar, and M. Gupta, ‘‘Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence,’’ Endocrine, vol. 78, no. 3, pp. 458–469, Oct. 2022, doi: 10.1007/s12020022-03215-4.

[2] (2023). World Health Organization. https://www.who.int/health-topics/obesity [Online]. Available:

[3] S. Rahman, M. Irfan, M. Raza, K. M. Ghori, S. Yaqoob, and M. Awais, ‘‘Performance analysis of boosting classifiers in recognizing activities of daily living,’’ Int. J. Environ. Res. Public Health, vol. 17, no. 3, p. 1082, Feb. 2020. [Online]. Available: https://www.mdpi.com/16604601/17/3/1082

[4] E. DeNicola, O. S. Aburizaiza, A. Siddique, H. Khwaja, and D. O. Carpenter, ‘‘Obesity and public health in the kingdom of Saudi Arabia,’’ Rev. Environ. Health, vol. 30, no. 3, pp. 191–205, 2015, doi: 10.1515/reveh-2015-0008.

[5] Z. A. Memish, C. El Bcheraoui, M. Tuffaha, M. Robinson, F. Daoud, S. Jaber, S. Mikhitarian, M. Al Saeedi, M. A. AlMazroa, A. H. Mokdad, and A. A. Al Rabeeah, ‘‘Obesity and associated factors—Kingdom of Saudi Arabia, 2013,’’ Preventing Chronic Disease, vol. 11, p. E174, Oct. 2014, doi: 10.5888/pcd11.140236. 13864

[6] F. A. Hamam, A. S. Eldalo, A. A. Alnofeie, W. Y. Alghamdi, S. S. Almutairi, and F. S. Badyan, ‘‘The association of eating habits and lifestyle with overweight and obesity among health sciences students in taif university, KSA,’’ J. Taibah Univ. Med. Sci., vol. 12, no. 3, pp. 249–260, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1658361216301494

[7] S.K.Keadle,R.McKinnon,B.I.Graubard,andR.P.Troiano,‘‘Prevalence and trends in physical activity among older adults in the United States: A comparison across three national surveys,’’ Preventive Med., vol. 89, pp. 37–43, Aug. 2016.

[8] A. C. Morrill and C. D. Chinn, ‘‘The obesity epidemic in the United States,’’ J. Public Health Policy, vol. 25, nos. 3–4,pp. 353–366, Dec. 2004, doi: 10.1057/palgrave.jphp.3190035.

[9] F. M. Palechor and A. D. L. H. Manotas, ‘‘Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, Peru and Mexico,’’ Data Brief, vol. 25, Aug. 2019, Art. no. 104344.

[10] G. Shao, ‘‘Comparison of prediction of obesity status based on different machine learning approaches with different factor quantities,’’ in Proc. Int. Conf. Biomed. Intell. Syst. (IC-BIS), Dec. 2022, p. 144.

[11] J. P. S. Quiroz, ‘‘Estimation of obesity levels based on dietary habits and condition physical using computational intelligence,’’ Informat. Med.

Unlocked, vol. 29, Jan. 2022, Art. no. 100901.
[12] I. G. S. M. Diayasa, M. Idhom, A. Fauzi, and A. T. Damaliana, ‘‘Stacking ensemble methods to predict obesity levels in adults,’’ in Proc. IEEE 8th Inf. Technol. Int. Seminar (ITIS), Oct. 2022, pp. 339–344.
[13] D.D.Solomon,S.Khan,S.Garg,G.Gupta,A.Almj ally,B.I.Alabduallah, H. S. Alsagri, M. M. Ibrahim, and A. M. A. Abdallah, ‘‘Hybrid majority voting: Prediction and classification model for obesity,’’ Diagnostics, vol. 13, no. 15, p. 2610, Aug. 2023.
[14] F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah, and M. T. Habib, ‘‘A machine learning approach for obesity risk prediction,’’ Current Res. Behav. Sci., vol. 2, Nov. 2021, Art. no. 100053.
[15] S. A. Thamrin, D. S. Arsyad, H. Kuswanto, A. Lawi, and S. Nasir, ‘‘Predicting obesity in adults using machine learning techniques: An analysis of Indonesian basic health research 2018,’’ Frontiers Nutrition, vol. 8, Jun. 2021, doi: 10.3389/fnut.2021.669155.
[16] Z. Zheng and K. Ruggiero, ‘‘Using machine learning to predict obesity in high school students,’’ in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), Nov. 2017, pp. 2132–2138.
[17] I. Machorro-Cano, G. Alor-Hernández, M. A. Paredes-Valverde, U. Ramos-Deonati, J. L. Sánchez-Cervantes, and L. Rodríguez-Mazahua, ‘‘PISIoT: A machine learning and IoT-based smart health platform for overweight and obesity control,’’ Appl. Sci., vol. 9, no. 15, p. 3037,
Jul. 2019.
[18] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, ‘‘An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI,’’ Sensors, vol. 22, no. 19, p. 7268, Sep. 2022.