Machine Learning Approaches for Obesity Classification and Prediction: An Analysis of Demographic, Lifestyle, and Health Factors

Joel Azu
Department of Science and Engineering
Solent University
Southampton, United Kingdom
ijoelazu@gmail.com

Raza Hasan
Department of Science and Engineering
Solent University
Southampton, United Kingdom
raza.hasan@solent.ac.uk

Shakeel Ahmad
Department of Science and Engineering
Solent University
Southampton, United Kingdom
shakeel.ahmad@solent.ac.uk

Salman Mahmood
Department of Computer Science
Nazeer Hussain University
Karachi, Pakistan
salman.mahmood@nhu.edu.pk

***Abstract*—The high level of obesity poses a serious public health problem across the globe, being a leading cause of various chronic diseases and substantial threats to the quality of people's lives. Given the ever-growing rates of obesity and its impact on individual well-being, the current study aims to explore numerous factors that drive obesity, using machine learning algorithms to solve classification tasks. By working on a rich dataset that contains a range of demographic, lifestyle, and health parameters, several classifiers were developed and tested, namely Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors. The resulting outcomes indicated that both Random Forest and Gradient Boosting algorithms were highly accurate in the classification of obesity, with 95.0% and 95.3% accuracy rates, respectively. The obtained results confirm the critical role of various machine learning approaches in understanding obesity and developing predictions for more focused intervention. This study offers considerable input into obesity epidemiology literature and demonstrates the utility of advanced analytical appraisal in public health.**

***Keywords*—Obesity Classification, Machine Learning, Ensemble Methods, Public Health, Predictive Analytics**

I. Introduction

Obesity is an epidemic public health problem that affects people of all ages, races and socioeconomic status. As defined by the World Health Organization [1], obesity is a disorder in which an abnormal or excessive fat accumulation may adversely affect a person's health. Obesity has many health consequences such as heart diseases, diabetes, cancer etc. [2].

Recent statistics from Public Health England [3] show that 63% of adults in the UK are classified as overweight (half moderate, half severe), which is a statistic, for reasons to be explored momentarily likely mirrored throughout many countries with similar 'developed world' socioeconomic/environmental systems. Among children it is no less alarming, with one in three 10-11 year olds overweight or obese by the time they leave primary school, and nearly a fifth of these falling into the obesity bracket. These statistics highlight the pressing demand for effective public health interventions to address growing disease rates of obesity.

Obesity is associated with numerous adverse health events, making the case for obesity to be targeted in both prevention and intervention efforts. In the realm of public discourse, it is also important to carry a nuanced blend between celebrating body positivity and recognizing that medical research has long recognized obesity as an incredibly serious health hazard. Educating people with information about the lifestyle causes of obesity is necessary to enable more informed choices and a healthier lifestyle.

In this study, we aimed at exploring the enigma of what determines obesity levels by making use of a rich dataset that covered different demographic, lifestyle and health conditions to serve in clearly evaluating its determinants beyond political preferences. This research seeks to elucidate the interplay of these factors and obesity levels using a comprehensive approach. Finally, this research aims to generate robust machine learning models to perform a high-quality classification of obesity adding important contributions for the field of epidemiology and public health with solutions addressing specific population segments.

II. Literature Review

Most recently, significant advancements have been shown in studies that implemented machine learning (ML) methods to predict obesity, with enhancement in the classifier accuracy and interpretability. The combination of Random Forest, Gradient Boosting and K-Nearest Neighbors algorithms perform exceptionally well in recognizing obesity status.

Gradient Boosting and K-Nearest Neighbors can predict obesity with high accuracy as reported by this study; however, their analysis was based on self-reported datasets and limited to using only BMI as the overly simplified measure of obesity which might lead to misclassification for athletic subjects due to ratio-based concern (i.e., it would be impossible even theoretically under most circumstances for a tall to reach overweight/obesity status compared to a short one in such an anthropometric study) [4].

Accuracy of up to 78% can be achieved by using ML techniques in predicting overweight or obesity [5]. Nonetheless, the study did not mention exploring broader assessments beyond the development of this predictive model.

A comparative analysis of ML techniques for obesity prediction was done by [6]. These methods can predict obesity (based on physical characteristics and dietary habits)

extremely well, however no information about their limitations was provided.

The authors conducted blood tests and checked blood pressure to assess risk factors of obesity. They found that screening tools differed in their accuracy by age-specific sex. Other common factors were rarely added, and the dataset imbalance issue was not mentioned [7].

A previous study [8] illustrated the importance of broad, multimodal datasets and more interpretable methods when predicting obesity during childhood or adolescence because existing models generally do not contain variables that come from multiple fields.

This study achieved good performance with low feature importance when predicting obesity in adults using logistic regression, however these models overlooked important factors like dietary quality and genetic structures that can be included by training on more extensive datasets [9].

Some other significant contributions are [10] and [11] that highlighted ML algorithms as promising tools for adult obesity prediction without stating their limitations.

III. METHODOLOGY

A. Dataset

This study uses a dataset from UC Irvine Machine Learning Repository [12] consisting of 2111 samples with 17 features. The key predictor variables are demographic (e.g., Gender, Age, Height, Weight), Lifestyle Factors (e.g., Family History of the disease, Frequency with which an individual consumes vegetable and does physical activity) and health markers.

B. Data Preparation

An initial check was conducted to ensure that no values were missing for any of the attributes. We found 24 duplicate records and removed them to avoid duplication of data. Furthermore, there was class imbalance in the target variable - Obesity Status. Random under sampling was used to correct this problem where WEKA equalized the dataset by discarding instances in majority class but keeping elements that are necessary. This process generated 1,904 records containing an equal number of events for each class (272 per event). Lastly, categorical variables were label-encoded to enable their use in vector calculations in subsequent modelling stages.

C. Exploratory Data Analysis (EDA)

EDA was also performed and many patterns within the dataset were explored. The first thing was to understand the numeric variables using basic descriptive statistics that gives an idea about central tendency and dispersion. Numerical features were presented with histograms and density plots to examine their distribution, these distributions were also assessed using the skewness (measure of data symmetry) and kurtosis as a rough measure of normality. Also, we built the correlation matrix between different variables to find which features are related and for feature selection. Lastly, feature importance was checked with multiple methodologies to identify which attributes were dominating the output variable.

D. Machine Learning Modeling

The performances of some classifiers are compared to predict levels of obesity in the test dataset, including Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). Stratified splits were used to split the dataset into training and testing subsets while maintaining obesity status class distribution across both sets. This is essential as it helps to preserve the representativeness of the minority classes, improving confidence in model performance. The training set was used to train the different models, where they learn from patterns and relationships present in a dataset. We evaluated the performance of each classifier with widely used metrics like accuracy, precision, recall, F1 score and considered confusion matrix as well ROC curve. Together, these metrics offer a complete view of how well the classifiers performed at predicting obesity statuses.

E. Model Selection

We have developed several machine learning models for the prediction of obesity levels. These models comprised the Random Forest classifier, Gradient Boosting Machine, Support Vector Machine, K-Nearest Neighbors, and Decision Tree. The performances of the models were done with regard to the performance metrics generally accepted, such as accuracy, precision, recall, F1 score, and the area under the ROC curve.

We have divided the dataset into 80% of training and 20% of testing subsets, using stratified sampling to keep the distribution of obesity categories in both sets. Accuracy was a basis for model selection, but also the overall balance of the other metrics: precision, recall, F1 score. The Gradient Boosting model had the highest accuracy at 95.3%, with the Random Forest coming closest at an accuracy of 95.0%. The models were chosen based on performance exhibited on the test dataset, balancing predictive accuracy and computational efficiency.

Thirdly, the confusion matrix for each classifier was plotted to investigate the robustness of the various classifiers in classifying the majority and minority obesity classes. Lastly, the ROC curves of all models were compared; the ensemble techniques, namely Random Forest and Gradient Boosting, indeed gave far better performance compared to other classifiers.

IV. RESULTS

Table I shows the overview of results from the classification models, where performance measures are described for each classifier used to predict obesity classes. The classifiers built were based on Decision Tree, Random Forest, Gradient Boosting, SVM and KNN algorithms.

TABLE I. MODEL EVALUATION METRICS

Classifier	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.927	0.929	0.927	0.926
Random Forest	0.95	0.953	0.95	0.95
Gradient Boosting	0.953	0.954	0.953	0.953
SVM	0.861	0.864	0.861	0.857

KNN	0.848	0.854	0.848	0.841

The performance of the models produced very good classification results; Gradient Boosting gave an accuracy of 95.3%, very closely followed by Random Forest at 95.0%. The outstanding performance is further reflected in the high precisions and recalls of both models, with Gradient Boosting giving an F1 score of 0.953, higher than that given by Random Forest, which gave an F1 score of 0.950. This confirms that the ensemble methods do really well for this classification as shown in Table I.

A. Confusion Matrices

The confusion matrices regarding individual classifiers are quite informative to evaluate their classification capabilities in terms of the true positive rate, false negative rate and so on. Fig. 1 is the confusion matrix for Decision Tree classifier; it shows how well-classified obesity statuses are. Similarly, the confusion matrix for the Random Forest model, shown in Fig. 2, presents a good performance with few misclassifications. The confusion matrix for the Gradient Boosting method is shown in Fig. 3 which provides us with an indication of how it can correctly classify between a variety of obesity categories. In contrast, the SVM confusion matrix however, which can be seen in Fig. 4 reveals a bigger number for false negative cases meaning that there are some obesity statuses it cannot identify correctly. Finally, Fig. 5 KNN Confusion Matrix, shows that classification was a bit difficult but without the recall being faulty as much in minority classes.

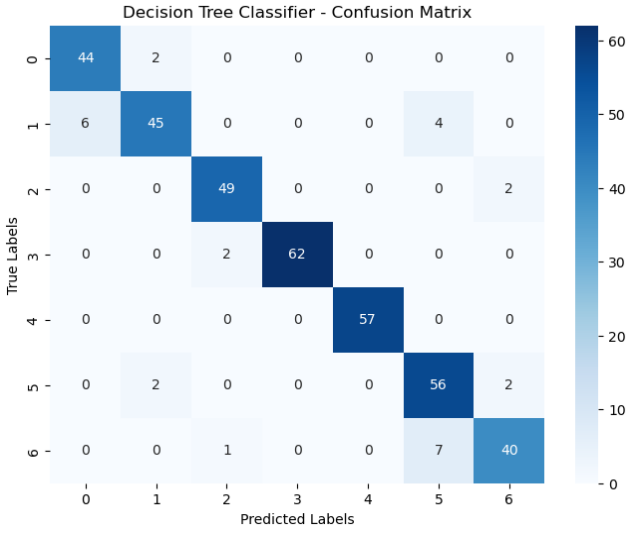

Fig. 1. Decision Tree Confusion Matrix

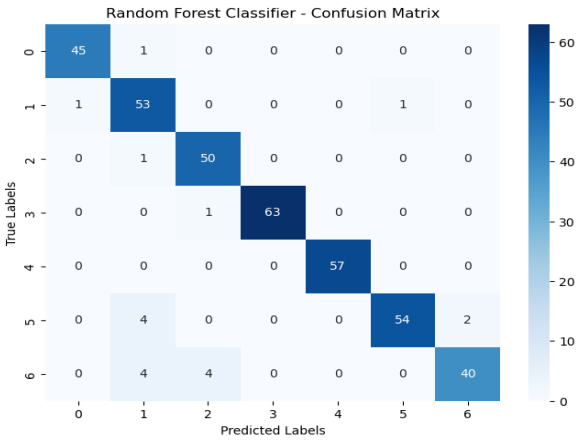

Fig. 2. Random Forest Confusion Matrix

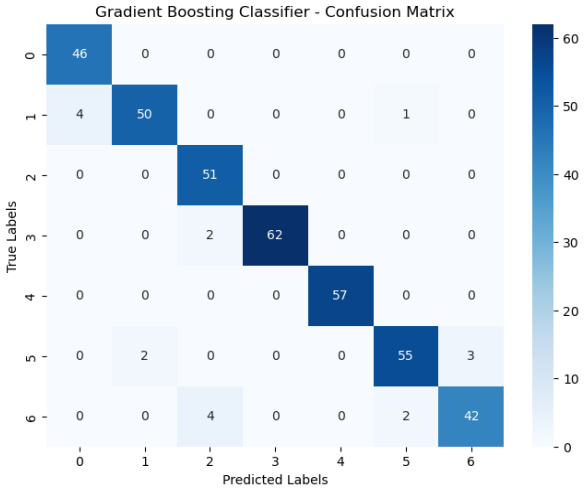

Fig. 3. Gradient Boosting Confusion Matrix

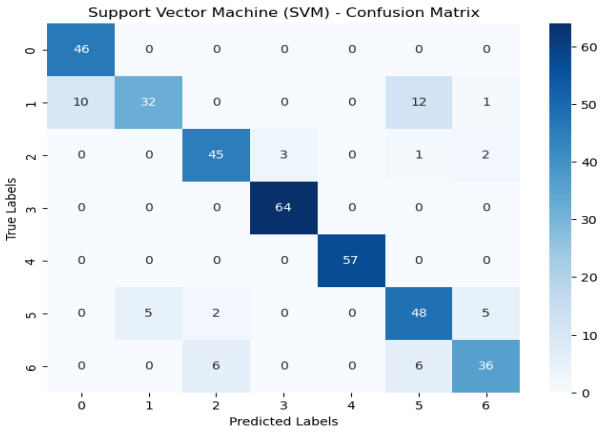

Fig. 4. SVM Confusion Matrix

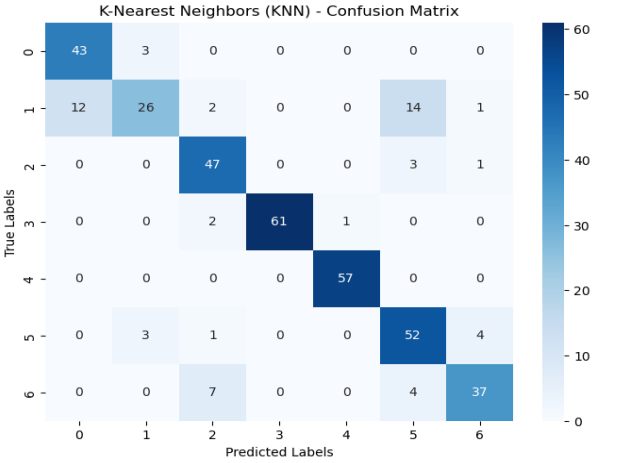

Fig. 5. KNN Confusion Matrix

The confusion matrices show that the RF and Gradient Boosting classifiers largely reduced misclassifications, especially in minority classes. SVM and KNN models had a higher rate of false negatives suggesting difficulty in accurately identifying some states of obesity.

B. Receiver Operating Characteristic (ROC) Curves

ROC curve for each classifier measures how good the classifier is at distinguishing between true positives vs false negatives across multiple thresholds. AUC (Area under the curve) values were determined to indicate how well models could classify easily into obesity categories. Fig. 6 is the ROC curve of the Decision Tree classifier, ROC curve for the Random Forest classifier is illustrated in Fig. 7. Gradient Boosting ROC curve is illustrated in Fig. 8. The ROC curves in the case of SVM and KNN classifier are illustrated in Fig. 9 and 10 respectively.

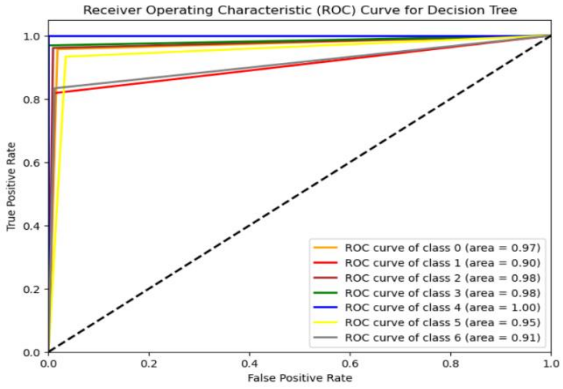

Fig. 6. Decision Tree ROC Curve

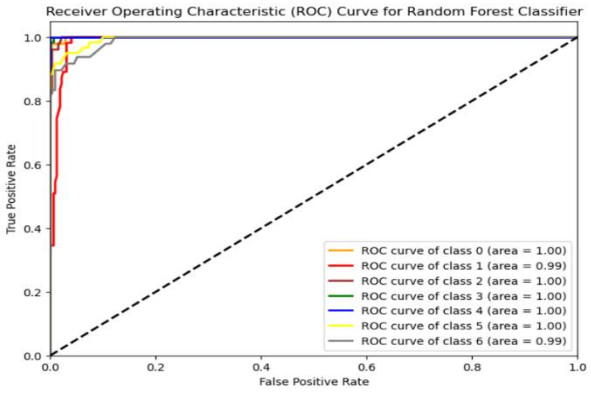

Fig. 7. Random Forest ROC Curve

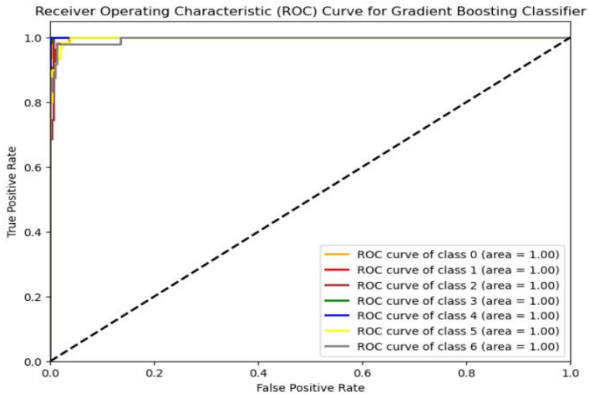

Fig. 8. Gradient Boosting ROC Curve

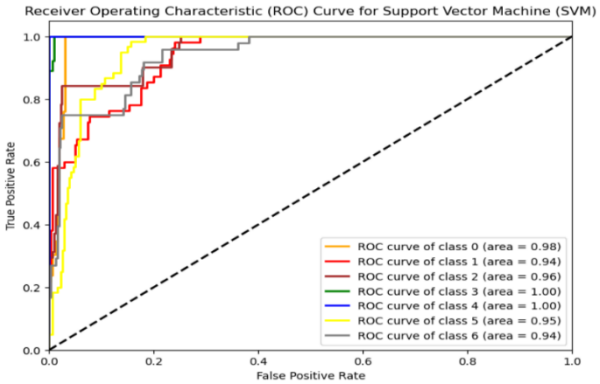

Fig. 9. SVM ROC Curve

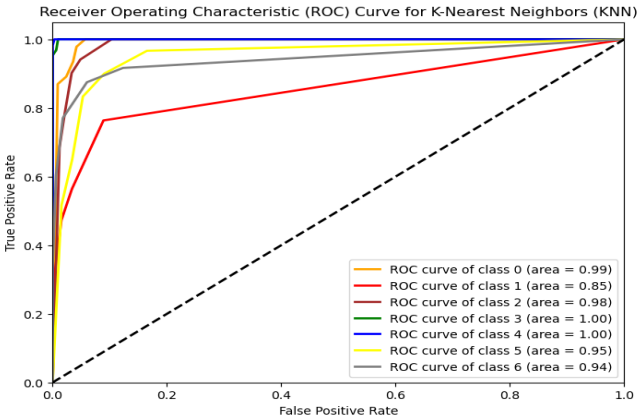

Fig. 10. KNN ROC Curve

ROC analysis indicates that the ensemble methods including Random Forest and Gradient Boosting achieved higher AUC measures in all experimentations. As a result, the experiment showed better performance with the two models, reinforcing that both Random Forest and Gradient Boosting have a better classification result than SVM and KNN. As illustrated, the best ensemble classifiers, especially Gradient Boosting and Random Forest, provide the best way of predicting the obesity level with higher accuracy levels and most of the metrics associated with it.

V. DISCUSSION

Results of this study emphasized that machine learning classifiers (such as ensemble methods) are remarkably

effective in predicting obesity levels according to lifestyle and demographic data. Random Forest and Gradient Boosting perform better than other classifiers, yielding the highest accuracy as well as F1 results. This is consistent with previous literature which has shown ensemble methods to be more resilient across a range of classification tasks as they are able to combine predictions from multiple models and therefore less prone to overfitting while improving generalizability.

Furthermore, we could observe that the performance of our models achieved results like those in prior studies. The performance of the Random Forest and Gradient Boosting classifiers awarded them the highest accuracy for the proposed dataset, which is in line with [13], [2] and [14]. As we can see in our evaluation, SVM and KNN showed the lowest results compared to the ensemble classifiers, which is consistent with findings from [15] and [16]. The highest performances of the ensemble methods, Random Forest and Gradient Boosting achieved better accuracy and F1 score, indicating stable performance for all classes of obesity. The algorithms like SVM and KNN, though acceptable in terms of results, still depicted low performances compared to the ensemble methods. For this task, the selection of ensemble methods may be desirable as they balance predictive performance with computational efficiency.

The confusion matrices (Fig. 1-5) demonstrate the acceptable performance of all classifiers, however using ensemble methods significantly improved our precision and recall scores especially for classes with lower support. This is of particular importance in obesity prediction, where inaccurate misclassification would carry public health implications.

The ROC curves (Fig. 6-10) demonstrate a superior lying setting compared to ensemble models based on AUC values. The higher the AUC values, the better it is in distinguishing different obesity categories which further supports these models as potential practical tools for public health settings.

However, the study noted some limitations. Reliance on a secondary dataset may have an influence in the generalizability of these results as it might not contain information representative of all population at risk of being obese. However, in future examinations, other complete and diverse datasets which contain samples from the general population should be used as this would make the prediction model as practical as possible.

VI. Conclusion

This study aimed to enhance our understanding of obesity epidemiology by exploring the effectiveness of machine learning classifiers in predicting obesity levels based on diverse lifestyle and demographic factors. The findings indicate that ensemble methods, specifically Random Forest and Gradient Boosting, outperformed other classifiers in terms of accuracy and F1 scores, demonstrating robust performance across all obesity categories. This reinforces the viability of these models for practical applications in public health.

While SVM and KNN provided satisfactory results, they were less effective compared to ensemble methods, highlighting the importance of model selection based on predictive performance and computational efficiency. The use of ROC curves further confirmed the superior discriminative ability of ensemble classifiers, making them ideal candidates for deployment in obesity prediction frameworks.

References

[1] World Health Organization, *"Obesity"* [Online]. Available: https://www.who.int/health-topics/obesity#tab=tab_1.

[2] M. Gupta *et al*, "Obesity Prediction with EHR Data: A deep learning approach with interpretable elements," *ACM Transactions on Computing for Healthcare*, vol. 3, *(3)*, pp. 1-19, 2022. Available: https://search.proquest.com/docview/2681440996. DOI: 10.1145/3506719.

[3] Public Health England, *"Obesity Profile"* [Online]. Available: https://fingertips.phe.org.uk/profile/national-child-measurement-programme.

[4] F. Musa, F. Basaky and O. E.O, "Obesity prediction using machine learning techniques," *Journal of Applied Artificial Intelligence*, vol. 3, *(1)*, pp. 24-33, 2022. Available: https://doi.org/10.48185/jaai.v3i1.470. DOI: 10.48185/jaai.v3i1.470.

[5] E. Rodríguez et al, "Machine learning techniques to predict overweight or obesity," in IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine.

[6] M. Dirik, "Application of machine learning techniques for obesity prediction: a comparative study," *Journal of Complexity in Health Sciences (Online)*, vol. 6, *(2)*, pp. 16-34, 2023. Available: https://www.extrica.com/article/23193/pdf. DOI: 10.21595/chs.2023.23193.

[7] J. Jeon, S. Lee and C. Oh, "Age-specific risk factors for the prediction of obesity using a machine learning approach," *Frontiers in Public Health*, vol. 10, pp. 998782, 2023. Available: https://www.ncbi.nlm.nih.gov/pubmed/36733276. DOI: 10.3389/fpubh.2022.998782.

[8] H. Siddiqui *et al*, "A Survey on Machine and Deep Learning Models for Childhood and Adolescent Obesity," *Access*, vol. 9, pp. 157337-157360, 2021. Available: https://ieeexplore.ieee.org/document/9627712. DOI: 10.1109/ACCESS.2021.3131128.

[9] S. A. Thamrin *et al*, "Predicting obesity in adults using machine learning techniques: An analysis of indonesian basic health research 2018," *Frontiers in Nutrition (Lausanne)*, vol. 8, pp. 669155, 2021. Available: https://search.proquest.com/docview/2549686517. DOI: 10.3389/fnut.2021.669155.

[10] K. N. Devi *et al*, "Machine Learning Based Adult Obesity Prediction," *Iccci*, pp. 1-5, 2022. Available: https://ieeexplore.ieee.org/document/9740995. DOI: 10.1109/ICCCI54379.2022.9740995.

[11] C. Suresh *et al*, "Obesity Prediction Based on Daily Lifestyle Habits and Other Factors Using Different Machine Learning Algorithms," *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*, pp. 397-407, 2022. Available: http://ebookcentral.proquest.com/lib/SITE_ID/reader.action?docID=6893917&ppg=411. DOI: 10.1007/978-981-16-7389-4_39.

[12] F. M. Palechor and A. d. l. H. Manotas, "Estimation of Obesity Levels Based On Eating Habits and Physical Condition," *Data in Brief*, vol. 25, pp. 104344, Aug 1, 2019.

[13] I. G. S. M. Diayasa *et al*, "Stacking ensemble methods to predict obesity levels in adults," in Oct 19, 2022, pp. 339-344.

[14] T. Khater *et al*, "Interpretable models for ML-based classification of obesity," in Aug 17, 2023, pp. 40-47.

[15] L. Wang *et al*, "Prioritization of multi-level risk factors for obesity," in Nov 2019, pp. 1065-1072.

[16] N. C. Pereira *et al*, "Obesity related disease prediction from healthcare communities using machine learning," in Jul 2019, pp. 1-7.