The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES

20(2): S2: 25-28, 2025

www.thebioscan.com

A REVIEW ON OBESITY PREDICTION USING MACHINE LEARNING TECHNIQUES

Dr. Kolluru Venkata Nagendra[1], Dr. Praveen B M[2]

PDF Research Scholar[1], Research Guide[2]

Department of Computer Science and Engineering[1 2],

Srinivasa University, Mangalore, Karnataka State, India.[1 2]

DOI: https://doi.org/10.63001/tbs.2025.v20.i02.S2.pp25-28

KEYWORDS
Detection of Obesity,
Machine Learning,
Logistic Regression,
Decision Tree,
KNN, SVM.

Received on:

12-02-2025

Accepted on:

10-03-2025

Published on:

08-04-2025

ABSTRACT

At present, protecting the community is crucial for addressing health issues, which can be done through medical research. Obesity has emerged as a global health crisis, posing a significant risk to the future. It ranks as one of the most prevalent health issues worldwide. Timely identification of a disease can assist both healthcare professionals and patients in taking action to reduce, if not completely eliminate, the underlying cause or in preventing the symptoms of the disease from worsening. Reviewing a patient's medical history is a common approach to diagnosing a disease; however, this process can be quite time-consuming when done manually and is often susceptible to errors and high costs. Thus, there is a compulsory to scientifically create a predictive model for the development of diseases using automated methods in today's world. This research highlights the capabilities of machine learning in tackling public health issues, laying the groundwork for future studies aimed at improving obesity prediction and prevention methods. Four machine learning algorithms were engaged: Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine. The results have been encouraging, with the Random Forest classifier got the highest accuracy at 96.93% among all.

INTRODUCTION

Globally, Obesity is one of the most frequent health issue linked to numerous illnesses, risks, and even mortality. It represents a significant health challenge on a worldwide scale, emerging as a potential danger for the future. As noted in [1] Mendoza, P. (2019), obesity has become a widespread and evolving global epidemic that has surged since the 1980s, raising serious health issues among adults, adolescents, and children alike. Furthermore, Eduado, D., Fabio, E., pointed out that the challenges associated with obesity are escalating swiftly, prompting new research that addresses obesity detection in both the young and the elderly.

According to this understanding, the WHO. (2021, June 23). *Obesity Related Diseases* [2] tells obesity as a rare or excessive collection of fat that can adversely affect health. Many individuals over the age of 16 are experiencing changes in their weight. This excess weight can result from a high consumption of starchy foods rich in fat, as well as a deficiency in physical activity. As noted in Guterrez, H. M. (2010). [3], obesity is a widespread health issue globally, impacting teenagers, children, and adults alike.

Obesity can be regarded as a multifaceted disease influenced by various factors, characterized by symptoms such as uncontrolled weight gain, primarily due to excessive consumption of energy-dense and fatty foods. Hernandez, J. (2011) [4] Research indicates that biological factors, including genetic predisposition, play a significant role in the development of different types of obesity, such as syndromic, monogenic, polygenic and leptin. Additionally, other risk factors, including dietary habits, social influences and psychological aspects, have been identified.

Globally, [5] represents a significant health challenge, as it is associated with chronic conditions such as cardiovascular diseases, diabetes, and cancer. Early detection of obesity has become a focal point in health initiatives. Nevertheless, extensive research over the years has revealed that effectively managing and preventing obesity is a complex endeavor. This complexity arises from our limited understanding of obesity and the intricate interplay of its various contributing factors, which encompass both environmental and biological elements.

This study focused on the specific objectives are as follows:

- To identify the perfect set of features for predicting obesity.
- To construct a model for predicting individual obesity using ML Techniques with a sourced data.
- To apply supervised ML methods to improve the accuracy of obesity level for future predictions.
- To evaluate the model's performance through established evaluation metrics.

II. REVIEW OF LITERATURE

This study Xiaolu, C., Shuo-yu - 2021[6] investigates the correlation between weight status and physical activity levels in humans, while also comparing various machine learning techniques with traditional statistical models for predicting obesity. The researchers utilized the National Health and Nutrition Examination Survey Dataset in their analysis, employing eleven distinct algorithms, including random subspace, logistic regression, decision table, Naïve Bayes, Radial Basis Function, K-nearest neighbor, classification via regression, J48, and multilayer perceptron. The evaluation metrics applied were the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC), with the random subspace classifier algorithm achieving the highest overall accuracy.

Employ machine learning techniques DeGregory, [7] such as linear and logistic regression, Artificial Neural Networks, Deep Learning, and Decision Tree analysis to forecast and classify obesity levels based on extensive datasets obtained from sensors, smartphones, and electronic health records. The findings indicate that machine learning offers advanced tools for predicting, classifying, and elucidating obesity-related risks and their consequences. A Zachary, J., Ward, M. P., [8] study aimed to forecast the obesity percentage among adults in the United States for the near future by applying multinomial regression to assess the prevalence of four BMI categories: under 25 as normal weight, 25 to under 30 as overweight, 30 to under 35 as moderate obesity, and above 35 as severe obesity. This analysis utilized a self-reported biased dataset from the Behavioral Risk Factor Surveillance System survey conducted between 1993-1994 and 1999-2016. The results suggested that by 2030, the adult obesity rate is projected to rise to 48.9%, with 24.2% classified as severely obese, and over 25% of adults in 25 states expected to be affected.

A study [9] Hagai, R., Smadar, S., Shiri, [2021] assessed the trend of BMI increases in children and developed a system to identify those at high risk of obesity before it reaches a critical level. Their findings revealed that the most significant rise in BMI occurs between the ages of 2 and 4, with accurate predictions being made for children aged 5 to 6 years. A Xuegin, P., [10] model was created to predict early childhood obesity using XGBoost, the ID3 decision tree model, and the Recurrent Neural Network (RNN) machine learning algorithm, utilizing electronic health record data. The results indicated that XGBoost achieved the highest area under the curve (AUC) value of 0.81 (p=0.001), outperforming all other models tested.

A logistic regression model was developed by Davila, P., DeGuzman [11] to assess the likelihood of BMI among children aged 2 to 17 in rural regions, utilizing publicly accessible datasets. The findings indicate that in smaller geographic areas, these estimates are crucial for formulating effective interventions and planning potential solutions, as the prevalence rates within census data range from 27% to 40%. In another study Manna, S., & Jewkes, A. M. [12] introduced a model employing fuzzy signatures to navigate the complexities associated with childhood obesity datasets, proposing a solution to address the risks linked to children's motor development and early onset of obesity.

Adnan, M. H. (2011). [13] presented an initial approach to predicting obesity, which was based on data gathered from primary sources, including parents, caregivers, and the children themselves. The authors aimed to identify risk factors such as obesity, parental education levels, children's habits, lifestyle, and environmental influences. Their proposed framework integrates a hybrid method combining decision trees and Naïve Bayes, referred to as NBTree. Adnan, M. H., Damanhoori, F. D., & Husain, W. (2010). [14] focused on predicting childhood obesity through data mining techniques. The objective of the proposed survey was to gather essential information regarding the obesity issue. The implementation involved models such as Neural Networks, Naïve Bayes, and Decision Trees.

The research conducted by Abdullah, F. S., Manan[2016] [15] examined the classification of obesity among sixth-grade children from two distinct districts in Malaysia. A classification technique was employed to analyze the collected data, utilizing machine learning models including Decision Trees, Support Vector Machines, Neural Networks, and Bayesian Networks. An article published by Tamara, M. D., [16] seeks to forecast obesity in children over the age of two by utilizing data gathered prior to their second birthday. The study examined six distinct machine learning techniques, including ID3, random tree, random forest, J48, Naïve Bayes, and Bayes, applied to the CHIKA dataset. The random forest method achieved an overall accuracy of 85%.

In another study Kapil, J., Niyati, B., & Prashant, S. R. (2018), [17] employed an ensemble machine learning approach to predict obesity in individuals, using attributes such as age and Body Mass Index (BMI), which is derived from weight and height. They implemented algorithms including the linear model, random forest, and partial least squares, achieving an accuracy of 89.68% with the random forest method. Research conducted by Brownlee, J. (2020), [18] focused on predicting future obesity levels in adults by applying Fitting Multinomial Regression to estimate the prevalence across four BMI categories: normal weight (less than 25), overweight (25 to less than 30), moderate obesity (30 to less than 35), and severe obesity (above 35), based on a self-reported biased dataset. A limitation of this study is its reliance on a single technique to address multiclass problems, along with the inherent bias in self-reported data concerning body measurements (height and weight). Furthermore, using only Body Mass Index as a metric can lead to inaccuracies, as it may misclassify athletic individuals with high BMIs due to muscle rather than excess body fat.

Therefore, it is crucial to incorporate additional features to develop a more sophisticated predictive model using machine learning techniques, rather than relying on simpler methods such as manual calculations or statistical approaches. This research aims to enhance an existing model to achieve greater efficiency and reduce human intervention in predicting obesity status, utilizing a native dataset with standardized components. In light of this, the present study explores the application of machine learning techniques to develop a predictive model for obesity in individuals, utilizing a locally sourced dataset from a healthcare diagnostic center that includes key features. This approach addresses the existing research gap.

III. EXISTING SYSTEM

In the initial stages, traditional and clinical methods are utilized to predict obesity, which can be a lengthy process as it necessitates the involvement of trained medical professionals to achieve a diagnosis. Often, complications arise from delayed interventions, as manual techniques are typically employed only after symptoms have become apparent in patients. Recently, numerous studies have explored the use of machine learning techniques for obesity prediction. These methods, adopted by various researchers, have yielded diverse results with varying degrees of accuracy, influenced by individual-related challenges. To diagnose obesity, healthcare providers conduct physical examinations and recommend specific tests. These assessments generally encompass gathering the patient's health history, performing a comprehensive physical examination (which includes measuring height, blood pressure, heart rate, and temperature), calculating the body mass index (BMI), measuring waist circumference, and evaluating lifestyle and genetic factors, as well as identifying any additional health issues. These procedures demand a significant level of expertise and considerable time investment.

a). Issues with the Current System

The current system faces several challenges:

i. Obesity diagnoses are often inaccurate due to the lack of expert evaluation for patients.

ii. The reliance on a single feature, specifically body measurements, may lead to incorrect predictions.

iii. The system is susceptible to low accuracy.

b). Proposed System Overview

The proposed model will be developed using the Python programming language, employing machine learning algorithms on a dataset obtained from a healthcare facility. This system aims to predict obesity in patients, utilizing a dataset that has been systematically automated to align with the objectives of this research.

The evolution of machine learning, a subset of artificial intelligence, has spurred various studies across multiple fields to enhance processes that were previously conducted manually. Once implemented, this model will assist healthcare professionals in predicting obesity in patients with reduced time commitment.

Additionally, it will facilitate the classification of obesity levels, thereby supporting timely interventions.

The system is expected to offer several benefits:

i. Enhanced reliability of diagnostic results.

ii. Minimization of the challenges associated with obesity diagnosis.

iii. Provision of a real-time classification system for obesity.

iv. Time savings and cost reductions associated with various medical tests.

IV. METHODS

Random Forest Classifier: The Random Forest Classifier is a versatile and user-friendly machine learning algorithm suitable for both classification and regression tasks. It operates by constructing an ensemble of decision trees, which are trained using the bagging method. This approach combines multiple learning models to enhance the overall performance of the model.

Decision Tree Classifier: The Decision Tree Classifier falls under the category of supervised learning algorithms and is employed to address regression and classification challenges. It utilizes a tree structure where each leaf node corresponds to specific attributes, while the internal nodes represent class labels. The entire training dataset serves as the root, and statistical methods are applied to categorize feature values and gather attributes for the internal nodes.

K-Nearest Neighbor: The K-Nearest Neighbor (KNN) algorithm is described as a classification technique that estimates the probability of a data point belonging to a particular group based on the proximity of other data points. KNN is a supervised machine learning algorithm applicable to both regression and classification tasks, though it is predominantly used for classification. This algorithm does not require training on the provided data and does not perform calculations; instead, it assesses the proximity of data points to determine their group affiliation.

Support Vector Machine: The Support Vector Machine (SVM) is a fundamental and adaptable algorithm in machine learning, recognized for its effectiveness in linear and nonlinear classification, regression, and outlier detection tasks. The SVM classifier is utilized for both classification and regression problems, with a primary focus on classification tasks. Due to its lower computational requirements and impressive accuracy, this algorithm is favored over other classification methods.

Evaluation: At this stage, the performance of our model will be assessed and verified to determine its correctness and accuracy. The evaluation will utilize the AUC value derived from our ROC curve to measure the model's accuracy.

Deployment: This is the concluding phase where the system is delivered to the end-user. It should be user-friendly and functional to ensure effective utilization.

V. RESULT

The evaluation of the prediction model was conducted by assessing the accuracy of each algorithm to determine its effectiveness in meeting the objectives. Presented below are the performance assessments for the four algorithms utilized: Random Forest Classifier, Decision Tree Classifier, K-Nearest Neighbor, and Support Vector Machine.

Table 1: Performance Evaluation of Models

Model	Accuracy
Random Forest	96.93%
Decision Tree	95.94%
K-Nearest Neighbor	95.68%
Support Vector Machine	97.17%

CONCLUSION

Diagnosing obesity presents challenges due to its multifaceted nature. There is a pressing need for enhanced diagnostic methods within the healthcare sector to minimize the associated risks and consequences. To accurately assess a patient's obesity status, physicians must conduct physical evaluations and analyze test results, which are often subject to the physician's simplification. This research aims to address the issues of diagnostic validity and time consumption by developing an obesity prediction model that offers a consistent diagnostic approach applicable to all genders. In this research, a prominent obesity prediction model has been created. The model utilizes machine learning techniques enforced in the Python programming language. The algorithms utilized are Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine. Among these, the Random Forest algorithm gives better performance of achieving accuracy 96.93%. This prediction model is strongly recommended for use in diagnostic centers, hospitals, clinics and the broader healthcare sectors.

REFERENCES

- [1] Eduado, D., Fabio, E., & Mendoza, P. (2019). Obesity Level Estimation Software based on Decision Tree. *Jornal of Computer Science*, 67-77.
- [2] WHO. (2021, June 23). *Obesity Related Diseases*. Retrieved from World Health Organisation: www.who.int
- [3] Guterrez, H. M. (2010). Diez problems de la poblacion de jalisco. *Una perspective sociodeografica*, 25-30.
- [4] Hernandez, J. (2011). Obesity and its causes. *International Journal for Medical Image*.
- [5] Del Cisne, P., & Zhingre, O. (2015). Factors that Influence Obesity. 10-13.
- [6] Xiaolu, C., Shuo-yu, L., Jin, L., Shiyong , L., Jun, Z., Peng, N., et al. (2021). Does Physical Activity Predict Obesity-A Machine Learning and Statistical Method-Based Analysis. *International Journal of environmental research and public Health*.
- [7] DeGregory, K. W., Patrick , K., DeSilvio, T., & Pleuss, J. D. (2018). A review of machine learning in obesity: machine learning in obesity reaserch. *ResearchGate*.
- [8] Zachary, J., Ward, M. P., Sara, N. B., Angie, L., Cradock, Jessical, L., et al. (2019). Projected US State Level Prevalence of Adult Obesity and Severe Obesity. *The New England Journal of Medicine*, 2440-2450.
- [9] Hagai, R., Smadar, S., Shiri, B.-H., Becca, F., Aron, W., & Eran, S. (2021). Prediction of Childhood Obesity from Nationwide Health Records. *The Journal of Pediatrics*, 132-140.
- [10] Xuegin, P., Christopher, B. F., Felice, L.-S., & Aaron, J. M. (2021). Prediction of Early Childhood Obesity with Machine Learning and Electronic Health Record Data. *International Journal of Medical Informatics*.
- [11] Davila, P., DeGuzman, C. M., Johnson, K., & Serban. (2015). Estimating prevalence of overweight or obese children and adolescents in small geographic areas uning publicly available data. *IEEE*.
- [12] Manna , S., & Jewkes, A. M. (2014). Understanding early childhood obesity risks: An empirical study using fuzzy signatures. *IEEE international confluence* (pp. 1333-1339). Beijing China : Xplore press.
- [13] Adnan, M. H. (2011). A framework for childhood obesity classifications and predictions using NBtree. *Proceedings of the 7th International Conference on Information Technology in Asia* (pp. 1-6). Kuching, Sarawak, Malaysia: IEEE Xplore Press.
- [14] Adnan, M. H., Damanhoori, F. D., & Husain, W. (2010). A survey on the usefulness of data mining for childhood obesity prediction. *Proceedings of the 8th AsiaPacific Symposium on Information and Telecommunication Technologies* (pp. 1-6). Kuching, Malaysia: IEEE Xplore Press.
- [15] Abdullah, F. S., Manan, N. S., Ahmad, A., Wafa, S. W., & Shahril, M. R. (2016). Data miningtechniques for classification of childhood obecity among year 6 school

children. *proceeding of the international conference on soft computing and data mining* (pp. 465-474). Springer: IEEE Xplore press.

- [16] Tamara, M. D., Mukhopadhyay, S., Aaron , C., & Stephen, M. D. (2015). Machine lerning techniques for prediction on early childhood obesity. *Researchgate*.
- [17] Kapil, J., Niyati, B., & Prashant, S. R. (2018). Obesity prediction using ensemble machine learning approaches . *Researchgate*.
- [18] Brownlee, J. (2020). Ensemble Learning. *Machine Learning Mastery*.
- [19] Donges, N. (2021). A Compete Guide to the Random Forest Algorithm. *Expert* Contributor Network.
- [20] www.geeksforgeeks.org. (2022). www.geeksforgeeks.org. Retrieved from www.geeksforgeeks.org.
- [21] Joby, A. (2021, July 19). learn.g2.com. Retrieved from G2 Learn Hub.
- [22] Dwivedi, R. (2021, January 29). Machine Learning. United States Artificial Intelligence Institute.