International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

Machine Learning Approaches For Obesity Level Classification

Shalini K[1], A.V. K. Shanthi[2], C. Shakila[3], N.Chamudeeswari [4]
[1]PG Scholar, Department of Data Science, SRM Institute of Science & Technology, Kattankulathur, Chengalpattu, Tamil Nadu. India.
[2]Professor, Department of Computer Science, Vel Tech RangaSangu Arts College, Avadi, Chennai, Tamil Nadu, India.
[3]Assistant Professor, Department of Computer Applications, Vel Tech RangaSankuArts College, Avadi, Chennai, Tamil Nadu, India
[4]Assistant Professor, Department of Computer Science, Vel Tech RangaSankuArts College, Avadi, Chennai, Tamil Nadu, India

Abstract: *Obesity has emerged as a global epidemic, posing significant public health challenges and increasing the risk of numerous chronic diseases. Accurate and timely classification of obesity levels is crucial for effective prevention, diagnosis, and personalized intervention strategies. Traditional methods, primarily relying on Body Mass Index (BMI), often fall short in capturing the complex physiological and lifestyle factors contributing to obesity. This paper explores the application of various Machine Learning (ML) approaches for the classification of obesity levels. We discuss diverse data sources, prominent ML algorithms, feature engineering techniques, and evaluation metrics pertinent to this domain. Furthermore, we highlight key challenges such as data imbalance, interpretability, and ethical considerations, alongside future directions for advancing ML in obesity management.*
Keywords: *Obesity, Machine Learning, Classification, BMI, Supervised Learning, Healthcare AI, Public Health.*

1. INTRODUCTION

Obesity is a complex, multifactorial chronic disease characterized by excessive body fat accumulation that poses a risk to health. According to the World Health Organization (WHO), worldwide obesity has nearly tripled since 1975, with more than 1.9 billion adults classified as overweight and over 650 million as obese in 2016. This global crisis is a major contributor to non-communicable diseases (NCDs) such as type 2 diabetes, cardiovascular diseases, certain cancers, and musculoskeletal disorders, placing immense strain on healthcare systems and reducing life expectancy and quality of life.

Accurate assessment and classification of obesity are foundational for targeted interventions. Traditionally, the Body Mass Index (BMI), calculated as weight (kg) divided by the square of height (m^2), has been the primary metric for classifying obesity levels. While widely adopted due to its simplicity, BMI has limitations; it does not differentiate between fat mass and muscle mass, nor does it account for fat distribution, age, gender, or ethnicity, leading to potential misclassifications. For instance, a highly muscular individual might be classified as overweight based on BMI, while an elderly person with sarcopenic obesity (low muscle mass, high fat mass) might fall into the "normal" BMI category.

The advent of machine learning offers a powerful paradigm shift in addressing these complexities. ML algorithms possess the unparalleled ability to identify intricate patterns and relationships within high-dimensional, heterogeneous datasets that may elude traditional statistical methods. By leveraging diverse data points – ranging from anthropometric measurements and clinical biomarkers to lifestyle behaviors, genetic predispositions, and environmental factors – ML can potentially provide a more nuanced, accurate, and personalized classification of obesity levels. This paper provides a comprehensive overview of how machine learning approaches are being utilized and can be further advanced for obesity level classification.

2. BACKGROUND: UNDERSTANDING OBESITY AND TRADITIONAL CLASSIFICATION

2.1. Defining Obesity and its Health Implications.

Obesity is medically defined as an abnormal or excessive fat accumulation that presents a risk to health. It is a disease with significant morbidity and mortality. Its pervasive impact extends to nearly every organ system, contributing to a cascade of comorbidities including:

Metabolic: Type 2 Diabetes, Insulin Resistance, Dyslipidemia, Metabolic Syndrome.
Cardiovascular: Hypertension, Coronary Artery Disease, Stroke, Heart Failure.

Respiratory: Sleep Apnea, Asthma.
Musculoskeletal: Osteoarthritis, Gout.
Gastrointestinal: Non-alcoholic Fatty Liver Disease (NAFLD), Gallstones.
Reproductive: Infertility, Polycystic Ovary Syndrome (PCOS).
Oncological: Increased risk of several cancers (e.g., colon, breast, prostate).
Psychological: Depression, Anxiety, poor body image.

2.2. Traditional Obesity Classification Methods.

The most widely used method for obesity classification is the Body Mass Index (BMI).
BMI Categories (WHO standards for adults):
Underweight: < 18.5
Normal weight: 18.5 – 24.9
Overweight: 25.0 – 29.9
Obese Class I: 30.0 – 34.9
Obese Class II: 35.0 – 39.9
Obese Class III (Morbid Obesity): ≥ 40.0
Note: Specific cut-offs can vary for Asian populations and children/adolescents.
Other traditional measures include:
Waist Circumference (WC): An indicator of abdominal obesity, which is strongly linked to metabolic syndrome and cardiovascular risk.
Waist-to-Hip Ratio (WHR): Another measure of fat distribution.
Body Fat Percentage: Measured by methods like Dual-energy X-ray Absorptiometry (DEXA), Bioelectrical Impedance Analysis (BIA), or skinfold thickness. These are more accurate but often less practical for routine screening.

2.3. Limitations of Traditional Methods

While practical, traditional methods, especially BMI, have inherent limitations:
Ignores Body Composition: BMI does not distinguish between fat mass and lean mass.
Does Not Account for Fat Distribution: Central adiposity (abdominal fat) is more metabolically harmful than peripheral adiposity, a distinction BMI fails to capture.
Variations by Age, Sex, Ethnicity: BMI thresholds may not be universally applicable without adjustment.
Limited Predictive Power for Metabolic Risk: Individuals with "normal" BMI can still be metabolically unhealthy (TOFI - Thin Outside, Fat Inside), and vice versa.
Lack of Dynamic Insight: They provide a static snapshot rather than continuous monitoring or insight into progression/regression.
These limitations underscore the need for more sophisticated approaches, which machine learning is well-positioned to offer.

3. Machine Learning Fundamentals for Classification

Machine learning is a subset of artificial intelligence that enables systems to learn from data, identify patterns, and make decisions or predictions with minimal human intervention. For obesity level classification, supervised learning is predominantly used, where the algorithm learns from a labeled dataset (i.e., examples where the obesity level is already known).

3.1. The ML Pipeline for Classification

A typical machine learning pipeline for classification tasks involves several stages:
Data Collection: Gathering relevant data from various sources.
Data Preprocessing: Cleaning the data (handling missing values, outliers), transforming it (normalization, scaling), and encoding categorical variables.
Feature Engineering: Selecting, combining, or creating new features from the raw data to improve model performance.
Model Selection: Choosing an appropriate ML algorithm based on the data characteristics and problem type.
Model Training: Feeding the preprocessed data to the chosen algorithm to learn patterns. This typically involves splitting data into training and validation/test sets to avoid overfitting.
Model Evaluation: Assessing the model's performance using specific metrics on unseen data.
Deployment: Integrating the trained model into a real-world application.

3.2. Types of Classification Tasks

For obesity, classification can be either:
Binary Classification: Distinguishing between two classes, e.g., "Obese" vs. "Non-Obese."
Multi-class Classification: Distinguishing between multiple predefined classes, e.g., "Normal Weight," "Overweight," "Obese Class I," "Obese Class II," "Obese Class III."

4. Machine Learning Approaches for Obesity Level Classification
The application of ML in obesity classification leverages a wide array of data types and sophisticated algorithms to build more robust and accurate predictive models.
4.1. Data Sources and Feature Engineering
The richness and diversity of input data are crucial for the success of ML models.
Anthropometric Data:
Core: Weight, Height, BMI (though BMI can also be an output label).
Advanced: Waist Circumference, Hip Circumference, Neck Circumference, Skinfold Thicknesses (triceps, subscapular), Body Fat Percentage (from BIA, DEXA).
Clinical and Laboratory Data:
Biomarkers: Blood glucose levels (fasting, HbA1c), insulin, lipid profile (cholesterol, triglycerides), liver enzymes, inflammatory markers (e.g., C-reactive protein), hormones (e.g., leptin, ghrelin, thyroid hormones).
Medical History: Presence of comorbidities (diabetes, hypertension, CVD), family history of obesity.
Pharmacological Data: Medications being taken (e.g., steroids, antipsychotics, which can induce weight gain).
Lifestyle and Behavioral Data:
Dietary Intake: Food frequency questionnaires, 24-hour recalls (calories, macronutrient distribution, food groups, eating patterns).
Physical Activity: Self-reported activity levels, wearable device data (steps, active minutes, energy expenditure).
Sleep Patterns: Sleep duration, quality, sleep disorders.
Smoking/Alcohol Consumption: Habits known to influence metabolic health.
Stress Levels: Psychosocial factors can influence eating behavior and weight.
Socio-Demographic Data: Age, Sex, Ethnicity, Socio-economic Status, Education Level, Geographic Location.
Genetic and Omics Data:
Genomics: Single Nucleotide Polymorphisms (SNPs) associated with obesity susceptibility (e.g., FTO gene).
Microbiomics: Gut microbiome composition and diversity, which is increasingly linked to metabolic health.
Metabolomics/Proteomics: Blood or urine metabolites/proteins that can indicate metabolic states.
Environmental Data: Access to healthy food, walkability of neighborhoods, pollution levels.
Feature Engineering: This critical step transforms raw data into a format that improves model performance.
Creating New Features: Ratios (e.g., Waist-to-Height Ratio), interaction terms (e.g., Age * Physical Activity).
Discretization/Binning: Grouping continuous variables into categories (e.g., age groups).
Normalization/Scaling: Ensuring features are on a comparable scale (e.g., Min-Max Scaling, Z-score standardization).
Feature Selection/Extraction: Reducing dimensionality by selecting the most relevant features (e.g., PCA, Recursive Feature Elimination, correlation analysis) to combat the "curse of dimensionality" and improve interpretability.
4.2. Common Machine Learning Algorithms for Classification
Various ML algorithms have been employed, each with its strengths and weaknesses:
Logistic Regression (LR):
Description: A linear model for binary or multi-class classification that estimates the probability of an instance belonging to a particular class.
Strengths: Simple, highly interpretable, computationally efficient, provides probability scores.
Application: Often used as a baseline model due to its simplicity and interpretability.

Support Vector Machines (SVM):
Description: Finds an optimal hyperplane that best separates different classes in the feature space, maximizing the margin between them. Can use kernel tricks to handle non-linear relationships.
Strengths: Effective in high-dimensional spaces, robust to overfitting, good for complex decision boundaries.
Application: Useful when the feature space is large and complex interactions are expected.
Decision Trees (DT):
Description: A tree-like model where each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class label.
Strengths: Easy to understand and interpret (white-box model), handles both numerical and categorical data, requires minimal data preprocessing.
Application: Provides interpretable rules that can be understood by clinicians.
Ensemble Methods (Boosting and Bagging):
Random Forest (RF):
Description: An ensemble method based on bagging, constructing multiple decision trees during training and outputting the mode of the classes.
Strengths: Reduces overfitting, high accuracy, robust to noise and missing data, provides feature importance scores.
Application: Widely used and often provides strong performance due to its robustness.
Gradient Boosting Machines (GBM - e.g., XGBoost, LightGBM, CatBoost):
Description: Builds trees sequentially, where each new tree corrects the errors of the previous ones.
Strengths: Highly accurate, handles complex non-linear relationships, excellent for structured tabular data.
Application: Often achieve state-of-the-art results in structured data classification challenges.
K-Nearest Neighbors (k-NN):
Description: A non-parametric, instance-based learning algorithm that classifies a data point based on the majority class among its 'k' nearest neighbors in the feature space.
Strengths: Simple to implement, no explicit training phase, effective for multi-modal data.
Application: Useful for exploratory analysis, though can be computationally expensive with large datasets.
Naive Bayes (NB):
Description: A probabilistic classifier based on Bayes' theorem, assuming strong independence between features.
Strengths: Simple, fast, and performs well with high-dimensional data, even with limited training data.
Application: Good for initial rapid prototyping or when computational resources are limited.
Neural Networks (NNs) / Deep Learning:
Description: Composed of layers of interconnected nodes (neurons) that can learn complex patterns through backpropagation. Deep learning refers to NNs with multiple hidden layers.
Strengths: Can model highly complex, non-linear relationships, excellent for learning from raw, unstructured data (e.g., image data from body scans, time-series data from wearables), can automatically learn features.
Application: While requiring large datasets, NNs can be powerful for integrating multi-modal data, including image data (e.g., from body scans) or complex sensor data from wearables to infer obesity.
4.3. Model Evaluation Metrics
For classification tasks, robust evaluation metrics are essential to understand a model's true performance, especially when dealing with imbalanced datasets (e.g., fewer morbidly obese individuals than normal weight).
Confusion Matrix: A table that summarizes the predictions of a classification model, showing True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).
Accuracy: (TP + TN) / Total. The proportion of correctly predicted instances.Can be misleading in imbalanced datasets.
Precision (Positive Predictive Value): TP / (TP + FP). The proportion of positive identifications that were actually correct.
Recall (Sensitivity/True Positive Rate): TP / (TP + FN). The proportion of actual positives that were correctly identified.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

F1-Score: The harmonic mean of Precision and Recall (2 * (Precision * Recall) / (Precision + Recall)). A balanced metric, especially useful for imbalanced classes.
Specificity (True Negative Rate): TN / (TN + FP). The proportion of actual negatives that were correctly identified.
Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):
ROC Curve: Plots the True Positive Rate against the False Positive Rate at various threshold settings.
AUC: The area under the ROC curve, representing the model's ability to distinguish between classes. A higher AUC indicates better performance.
Kappa Statistic (Cohen's Kappa): Measures the agreement between the predicted and actual classifications, accounting for chance agreement.
Cross-Validation: Techniques like k-fold cross-validation are crucial to ensure model generalizability by training and testing the model on different subsets of the data.

5. CHALLENGES AND CONSIDERATIONS

Despite the promise, several challenges must be addressed for successful and ethical deployment of ML in obesity classification:
Data Availability and Quality:
Limited Access: High-quality, diverse, and large datasets (especially those combining clinical, genomic, and lifestyle data) are often scarce due to privacy regulations and data silos.
Missing Data: Incomplete records are common in clinical datasets.
Data Bias: Datasets might not be representative of the global population, leading to models that perform poorly on underrepresented groups.
Data Imbalance: Some obesity classes (e.g., Class II or III obesity, or underweight) are less prevalent than others, leading to models that are biased towards the majority class.
Interpretability and Explainability (XAI):
Many powerful ML models (e.g., Deep Learning, complex ensemble methods) operate as "black boxes," making it difficult to understand why a particular classification was made.
In healthcare, interpretability is crucial for clinician trust, identifying potential biases, and providing actionable insights for patients. Explainable AI (XAI) techniques are vital here.
Causality vs. Correlation:
ML models primarily identify correlations within data. Establishing causality (e.g., which specific factors directly cause obesity or its progression) often requires further experimentation and domain expertise.
Ethical and Privacy Concerns:
Health data is highly sensitive. Ensuring data privacy, security, and ethical use (e.g., preventing discrimination based on predicted obesity risk) is paramount.
Models must be developed with fairness in mind, avoiding perpetuation or amplification of existing health disparities.
Generalizability and Real-World Deployment:
A model trained on one population or healthcare system might not perform equally well when deployed in a different context due to variations in demographics, lifestyle, and healthcare practices.
Integration into existing clinical workflows requires careful planning and validation.
Dynamic Nature of Obesity:
Obesity is not static; it involves progression, remission, and the impact of interventions. Static classification models may not capture this dynamic nature effectively.

6. FUTURE DIRECTIONS

The field of machine learning for obesity classification is rapidly evolving, with several promising avenues for future research and development:
Integration of Multi-Modal and Longitudinal Data: Combining diverse data sources (clinical, genetic, microbiome, lifestyle, environmental, wearable sensor data) will create richer datasets for more comprehensive and accurate models. Longitudinal data will allow for predicting progression or response to interventions.
Explainable AI (XAI) for Clinical Decision Support: Developing and integrating XAI techniques (e.g., SHAP, LIME) to provide transparent and actionable insights from "black box" models. This will foster trust and facilitate acceptance by clinicians and patients.

Personalized Risk Prediction and Intervention: Moving beyond mere classification to personalized risk profiles and tailored recommendations. ML can identify granular subgroups of patients who respond differently to various interventions, paving the way for precision nutrition and exercise prescriptions.
Reinforcement Learning for Dynamic Interventions: Exploring reinforcement learning to develop adaptive, real-time intervention strategies that learn from a patient's responses and adjust recommendations dynamically (e.g., personalized nudges for diet or activity based on current behavior and physiological state).
Federated Learning: To address data privacy concerns and leverage distributed datasets, federated learning can enable models to be trained across multiple healthcare institutions without centralizing sensitive patient data.
Leveraging Unstructured Data: Advanced NLP techniques can extract valuable information from clinical notes and patient narratives, while computer vision can analyze body composition from medical imaging or even standard photographs.
Internet of Medical Things (IoMT) and Wearables: Continuous monitoring of physiological parameters (heart rate, activity levels, sleep patterns) and behavioral data from IoMT devices and wearables can provide real-time, high-resolution data streams for dynamic prediction and intervention.

7. CONCLUSION

Obesity is a multifaceted global health challenge that necessitates innovative approaches for accurate assessment and effective management. Machine learning offers a transformative potential to move beyond the limitations of traditional BMI-centric classification. By integrating diverse data sources and employing sophisticated algorithms, ML models can provide more precise, nuanced, and personalized obesity level classifications, considering the complex interplay of genetic, physiological, behavioral, and environmental factors.
While significant challenges remain, particularly concerning data availability, interpretability, and ethical considerations, ongoing advancements in data science, computational power, and AI research offer promising solutions. The future of obesity management will likely involve a collaborative synergy between machine learning and clinical expertise, leading to earlier detection, more accurate risk stratification, and highly personalized interventions that can ultimately improve public health outcomes on a global scale. The journey towards fully integrated and intelligent obesity classification systems is challenging but holds immense promise for revolutionizing the fight against this pervasive chronic disease.

REFERENCES

[1] World Health Organization. (2021). Obesity and overweight. Retrieved from [WHO website]
[2] GBD 2021 Risk Factor Collaborators. “Global Burden of 88 Risk Factors in 204 Countries and Territories, 1990–2021: a systematic analysis for the Global Burden of Disease study 2021”. *Lancet.* 2024; 403:2162-2203.
[3] Okunogbe et al., “Economic Impacts of Overweight and Obesity.” 2nd Edition with Estimates for 161 Countries.World Obesity Federation, 2022.
[4] RajdeepKaur, Rakesh Kumar, Meenu Gupta “Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence”
,https://www.researchgate.net/publication/364319461_Predicting_risk_of_obesity_and_meal_planning_to_reduce_the_obese_in_adulthood_using_artificial_intelligence
[5]American Medical Association. (2013). Resolution 420 (A-13) - Recognition of Obesity as a Disease.
[6] Haslam, D. W., & James, W. P. (2005). Obesity. The Lancet, 366(9492), 1197-1209.
[7] Kaggle Datasets on Obesity Prediction/Classification.(Numerous user-contributed datasets and notebooks).
[8] Chen, T., &Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[9] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions.Advances in Neural Information Processing Systems, 30.
[10] Rieke, N., Hancox, J., Li, W., Liu, H., Plesser, M., Ward, G., ...&Kather, J. N. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 1-8.
[11] Specific research papers implementing ML algorithms for obesity classification (e.g., studies using SVM, RF, DNN on various datasets).
[12] SUSMITA BAG AND ANAND ANBARASU, OBESITY: A CRITICAL REVIEW, International Journal of Pharma and Bio Sciences, Volume 2 Issue 4, 2011 (October - December), Pages:582-592.
[13] MahmoodSafaeia ,Elankovan A. Sundararajan a,* , MahaDrissb,c , WadiiBoulilab,c , AzrulhizamShapi'i d , ‘A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity’, Computers in Biology and Medicine.
https://doi.org/10.1016/j.compbiomed.2021.104754. journal homepage: www.elsevier.com/locate/compbiomed,

[14] AnarTaghiyev, AdemAlpaslanAltun, SonaCaglar, A Hybrid Approach based on Machine Learning to Identify the Causes of Obesity. Technical Report 2, 2020.
[15] Isaac Machorro-Cano, GinerAlor-Herna´ndez, Mario Andre´s Paredes-Valverde,Uriel Ramos-Deonati, Jose´ Luis S´anchez-Cervantes, Lisbeth Rodríguez-Mazahua, PISIoT: a machine learning and IoT-based smart health platform for overweight and obesity control, Appl. Sci. 9 (15) (2019) 3037.
[16] Jiao Wang, Zhiqiang Deng, Modeling and prediction of oyster norovirus outbreaks along Gulf of Mexico Coast, Environ. Health Perspect. 124 (5) (2016) 627–633.
[17] Yang CY, Peng CY, Liu YC, Chen WZ, Chiou WK. Surface anthropometric indices in obesity-related metabolic diseases and cancers. Chang Gung Med J, 34(1):1-22, (2011).
[18]ShabanaTharkar, Vijay Viswanathan,"Effect of obesity on cardiovascular risk factors in urban population in South India", January 2010, Heart Asia 2(1):145-149, DOI: 10.1136/ha.2009.000950.