Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2025 (43) 4** : 1–17.
ISSN-0037-8712
doi:10.5269/bspm.78576

Obesity Risk Prediction Using Fusion Ensembling Methods

Sai Charan Medaramatla, Shekshavali Pattan, Sai Harshavardhan Srungavarapu, Voddelli Srilakshmi*

ABSTRACT: Obesity is a growing global health concern, and early risk prediction plays a vital role in enabling timely intervention. This study presents a novel hybrid ensemble model for obesity risk prediction using a dataset containing 2,111 samples and 17 features related to demographic, anthropometric, and lifestyle factors. The proposed model integrates Bagging through Random Forest and Boosting techniques using XGBoost, LightGBM, and CatBoost. These models are trained independently, and their outputs are combined using a simple averaging strategy to enhance prediction accuracy while reducing overfitting and bias.

The hybrid model achieved an accuracy of 97.85%, significantly outperforming traditional models such as Logistic Regression (86%), KNN (80%), and SVM (88%), as well as standalone ensemble methods. Comprehensive preprocessing ensured balanced classes and preserved meaningful outliers, and hyperparameter tuning was employed to optimize each base model. Additionally, feature importance analysis revealed key predictors, including meal frequency, physical activity, and water intake. This model demonstrates strong potential for use in clinical decision support systems and public health monitoring tools. The results highlight the effectiveness of fusion ensembling in handling complex classification problems in healthcare datasets.

Key Words: Obesity, risk prediction, ensemble techniques, random forest, XG Boost.

Contents

* Corresponding author.
2010 *Mathematics Subject Classification:* 68T05, 92C50.
Submitted August 22, 2025. Published December 20, 2025

1. Introduction

One of the primary health concerns these days is obesity, especially in children. Obesity and malnutrition are goals of the UN Decade of Action on Nutrition [2]. The change in the lifestyle of people is bringing these unwanted diseases, and the more common availability of high-calorie and processed food is the main reason for the increasing obesity rates. The lifestyle changes are affecting people of all ages and genders with obesity, the diseases are also chronic and are known to be very hard to cure [11]. The major types of chronic diseases that are caused by obesity are diabetes, cardiovascular diseases and cancer. Many people are also trying to cure their obesity through dietary changes. It is the most common and traditional method to cure obesity. The changes in lifestyle and dietary intake usually show many chances to reduce obesity in a person. The person suffering from obesity should maintain a balanced diet that is rich in proteins, fruits, vegetables, and grains. The lifestyle change should also include daily physical exercise, simple activities like walking and running are the activities most people prefer and see changes in [4]. The people who work out usually do strength training as that helps in building muscle mass and burning calories.

Another traditional method is controlling obesity and trying to bring changes in obesity-causing habits is behavioural therapy. To become a healthy person many go for cognitive behavioural therapy which helps the person to easily identify the habits that are harming them and tries to reduce those habits. The self-monitoring habits like maintaining records of the foods that are being eaten also help in controlling unhealthy food which will help in controlling obesity or even the risk of obesity occurrence. Spending more time with people with similar goals will also help us to learn more information about our condition and new ways to cope with the situation. Even though there are many methods to help in controlling obesity many people don't have access to the information regarding them [10]. Avoiding the concerns and ignoring the obesity condition is considered a huge risk. One of the most commonly affected chronic diseases contains the two types of diabetes. Obesity causes a lot of body fat in various parts, even though it might not be a problem in an early stage later on it becomes very difficult even for simple tasks such as lifting a pen off the ground. The fat growth in the body increases insulin resistance and the blood sugar levels will be hard to regulate. The excessive weight that a person grows due to obesity will put pressure on the heart leading to diseases like hypertension, coronary artery disease, and stroke. cancer is also one of the common chronic diseases people with obesity might get [1]. The other most common suffering many face due to obesity is depression, various mental health issues arise once obesity starts. Discrimination in society is one of the main reasons obese people suffer from mental health issues. Physical limitations and reduced mobility are also a part of obesity that many people wish to avoid. There are many more various physical and mental health conditions [9] that occur due to obesity, which is why it is never too soon to predict obesity occurrence.

These days AI and ML are being used in every field, they can also be used to forecast the risk of obesity in a person. Most of the existing models are using the traditional or basic machine learning algorithms. Regression analysis is a frequently used algorithm, many people are using it due to its exceptional accuracy rates. The popular ML algorithms in this field are Decision Tree, Random Forest, and SVM [6]. these algorithms are useful even in cases of large data. The prediction models work by analyzing the patterns and relations within the chosen dataset. More information and comparison of the existing methods and methodologies are explained in the related work section. The proposed model

The following sections detail the three key contributions of this research.

- The study develops a hybrid model using various ensemble techniques to forecast the risk of obesity.
- The developed hybrid model achieved an accuracy of 0.97.
- The model developed beats all the existing algorithms performance with much less complexity.

The following sections will delve deeper into our proposed approach. Section 2 regards the popular existing methods in predicting the risk of obesity in recent years; Section 3 contains information about the methods and proposed methodology; Section 4 is the discussion about the results obtained and comparison of the existing results with the proposed model; Section 5 contains the conclusion and future scope.

2. Related Work

[12] were concerned about the increasing obesity cases in the United States and developed a model that forecasts the chance or risk of obesity in children. The model developed by the authors uses basic factors like BMI, gestational age, and gender of the child to predict the risk factor. The dataset used in this work to build the model is divided into three categories: numerous well-child visits, lone well-child visits, and random well-child visits in the first group. Regression analysis and ML-based classification methods were used by the scientists to calculate the risk factor. On average, the model obtained an accuracy of 89%. The limitation of this study that is mentioned by the authors is that the dataset is considered very small in size containing data of less than 250 children.

[7] thought that obesity is one of the most complex biological problems, as it does not have one specific reason. When they observed the increasing rate of obesity in Korea, the authors felt the need to conduct a study in this sector and help society predict the occurrence and obesity. To train the model, the authors used a publicly available dataset of blood tests and blood pressure. During the model evaluation, the developed model is compared to contrast with existing models with five other machine learning classifier models. The results that were obtained are considered accurate as they were also based on age and gender, the highest accuracy obtained was 70% and it belongs to the class of 19-39 age group.

[16] felt that the factors that are causing obesity are changing daily and there has to be some advanced research on this topic to accurately estimate the obesity levels. The two main factors that were considered for to experiment are physical activity and eating habits. Some of the classification algorithms that were used to train the model are Chi-Square, and F-Classify, these help in obtaining the crucial information from the dataset easily. To evaluate the developed model's performance, it is compared with trained neural networks of various feature sets. To make it more perfect and faster, the hyperparameters are updated using the Bayesian techniques. The best accuracy achieved by the developed model is 93.06% and F-Classify showed the best performance among other algorithms, identifying the essential features to accurately estimate obesity levels.

[15] looked into a novel approach to utilizing machine learning to predict and categorize obesity. A serious health issue, obesity is connected to numerous other illnesses. A combination of heredity, bodily processes, environment, food, and exercise is to blame. Modern measures such as BMI can be inaccurate, particularly for individuals with large frames. Three distinct machine learning algorithms are combined in the model developed in this study. Using available data, they experimented with several methods and discovered that, with a 97.16% accuracy, their combined model performed the best. This implies that the new model for diagnosing obesity is superior to earlier techniques.

[5] study employed machine learning algorithms that consider physiological and genetic to predict the risk of obesity. The study utilizes a large dataset containing demographic information, physical activity levels, dietary habits, and genetic markers. Various ML methods namely logistic regression, random forests, SVMs, neural networks, and decision trees are applied and evaluated using standard performance metrics. To identify the most significant predictors of obesity, the paper emphasizes thorough feature selection and engineering techniques. Moreover, ensemble methods like gradient boosting and stacking are studied to increase prediction accuracy by leveraging the benefits of several models. The primary objective is to develop reliable prediction models for early detection and intervention planning among individuals at higher risk of obesity. This research aims to advance personalized healthcare strategies in tackling global health challenges associated with obesity-related diseases.

[3] observed the trends in the obesity of children in China and decided to conduct a study to predict the epidemic in seven to eighteen-year-olds. The regression function that is used to predict is polynomial regression, which is best known for its flexibility and non-linear relationships. Using advanced predictive modelling techniques, the study forecasts future trajectories of obesity prevalence, providing critical insights essential for shaping public health strategies and policies. The authors concluded that based on the previous years' data, there will be a rapid growth percent of people that will become obese.

The research paper by [8] introduces DeepHealthNet, an innovative system developed to predict adolescent obesity using deep learning methods. Utilizing a deep neural network architecture, DeepHealthNet processes a wide range of health data inputs to forecast the likelihood of obesity among adolescents. By integrating diverse factors such as demographic details, dietary patterns, physical activity levels, and medical histories, DeepHealthNet aims to improve the precision and dependability of obesity predictions.

Title	Method	Accuracy
Pritom Kumar Mandal et al.	Regression analysis and machine learning-based classification algorithms	89%
Junhwi Jeon et al.	Logistic Regression, Random Forest, SVM	70%
Fatma Hilal Yagin et al.	Chi-Square, F-Classify, neural networks	93.06%
Dhalak Daniel Solomon et al.	Gradient booster, XGB Classifier, MLP	97.16%
Faria Ferdowsy et al.	Logistic Regression, Random Forest, SVMs, neural networks, Decision Trees	97.09%
Dong YH et al.	Polynomial regression	Varies
Ji Hoon Jeong et al.	Deep learning with deep neural network architecture	90%
Casimiro Aday Curbelo Montanez et al.	Random forests, support vector machines, neural networks	88.4%
Balbir Singh et al.	K Nearest Neighbour algorithm, SMOTE	90%

Table 1: Comparison of Methods and Accuracies

The study emphasizes the efficacy of deep learning in managing complex, multidimensional health data and highlights its potential in advancing personalized healthcare approaches aimed at combating obesity in younger populations. The average accuracy of the developed model is shown to be 90%.

This study conducted by [13] explores the prediction of obesity using ML approaches and publically available genetic data profiles. Its main objective is to use computational techniques to analyze large genomic datasets extracted from repositories to identify specific genetic markers and patterns associated with the risk of obesity. This study thoroughly evaluates a variety of ML algorithms designed for certain objectives including feature selection and model construction. This research aims to use advanced analytics to integrate genomic data and improve our understanding of how hereditary factors affect the risk of obesity. This approach seeks to increase knowledge of the underlying genetics of obesity to increase the accuracy of tailored risk assessment techniques. The ultimate objective of the study is to advance precision medicine by developing robust predictive models that enable tailored treatments and early detection of obesity by utilizing information gleaned from genetic predispositions.

[14] identified obesity as one of the most concerning problems and developed a model that is based on ML to forecast a young person's risk of obesity or overweight. The dataset that is used to develop the proposed model is generated by the UK's millennium cohort study. The body mass indexes of children ages 3, 5, 7, and 11 were considered and based on their obesity level three classes were divided. However, there were a lot of imbalances in the classes, so the authors used the SMOTE technique to balance the classes. The K Nearest Neighbour algorithm is used to develop the proposed model. The results were compared with various traditional algorithms such as Random Forest, SVM, and MLP. The Multi-Layer Perceptron showed the best results with an accuracy of 90% for the balanced classes but KNN showed the best results even with unbalanced classes.

Even though many of the existing methods give good accuracy percentages, there are limitations in many experiments, one of the most common limitations of these existing methods is using traditional algorithms, the usage of hybrid or modern algorithms might produce even better accuracy. Another common limitation is the dataset, most of the datasets are either small or have a lot of noise these require a lot of data preprocessing and augmentation techniques, if the dataset is self-produced or self-collected, noise in the data might decrease leading to producing a more efficient model.

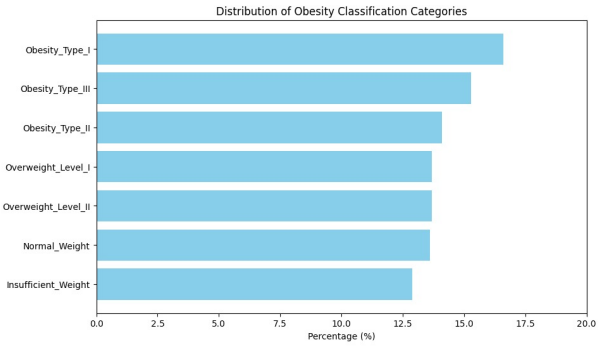

Figure 1: Distribution of classes in dataset

3. Methodology

3.1. Dataset Description

The "Obesity Data Set: Raw and Synthetic" has 2,111 entries and 17 variables, providing a complete spectrum of demographic, anthropometric, and lifestyle parameters linked to obesity. Height, weight, age, gender, and dietary habits, such as the frequency of high-calorie meals, the amount of vegetables consumed, and meal planning, are significant variables. Information is also given about lifestyle factors such as alcohol consumption, levels of physical activity, smoking status, and the amount of time spent using technology. The dataset includes an obesity classification system that categorises individuals into various levels, from normal weight to different degrees of overweight and obesity. The classes in the dataset are Class_1: Insufficient_weight, Class_2: Normal_Weight, Class_3: Obesity_Type_I, Class_4: Obesity_Type_II, Class_5: Obesity_Type_III, Class_6: Overweight_Level_I, Class_7: Overweight_Level_II. With no missing values, this dataset offers a robust foundation for exploring the relationships between lifestyle choices and obesity, making it an invaluable resource for public health research and the development of predictive models.

3.2. Data Preprocessing

Data preprocessing for the "Obesity Data Set: Raw and Synthetic" involved several critical steps to ensure the dataset was prepared for analysis. Initially, all records were inspected for completeness, confirming that no missing values existed across the 17 attributes. To make categorical data easier to use in machine learning models, such as gender, eating habits, and obesity levels, one-hot encoding was applied to the data. Continuous variables, including age, height, weight, and physical activity levels, were standardised to a common scale to prevent any one variable from disproportionately influencing the analysis. Outliers were identified and assessed, but due to the synthetic nature of the data, these were retained unless they were found to skew the results significantly. Additionally, correlations between variables were examined to identify and handle any multicollinearity, ensuring that the input features were independent and contributing uniquely to the model. This thorough preprocessing ensured that the dataset was clean, well-structured, and suitable for robust statistical analysis and machine learning applications.

The dataset appears to be fairly balanced across the different obesity classification categories. The distribution of classes is shown in Fig. 1

Each class represents a similar proportion of the dataset, with the smallest category (Insufficient_Weight) comprising about 12.9% and the largest (Obesity_Type_I) at 16.6%. The dataset appears to be balanced based on this reasonably even distribution, which is good for training ML models because it decreases the likelihood of bias towards any one class.

Although some outliers were identified, they were retained due to the synthetic nature of the dataset. Removing them might have distorted the data's diversity, which is essential for training a model capable

of handling real-world variability.

3.3. Model Description

3.3.1. Logistic Regression. Logistic regression is a statistical method that is used to predict the probability of a binary event occurring. By converting input variables onto a scale between 0 and 1, logarithmic regression ensures that the projected probabilities fall within this range. This makes it ideal for tasks like dividing data into two categories. By training on a dataset with known outcomes, the model discovers the relationship between input factors and the projected probability. It can be trained with fresh, unseen data to forecast the probability of an event.

3.3.2. K-Nearest Neighbors (KNN). K-Nearest Neighbors (KNN) is a simple yet effective ML algorithm used for classification tasks. KNN's basic tenet is that instances with comparable properties are probably members of the same class. By determining a new data point's k nearest neighbours from the training set, a KNN classifies it. The new point is subsequently allocated to the class of most of these neighbours. To maximize the model's output, the hyperparameter k, which controls the number of neighbours taken into account, can be adjusted. KNN doesn't assume any specific shape or pattern for the data, making it flexible for various datasets. This is because it's a non-parametric technique, meaning it doesn't have fixed parameters that need to be set beforehand. As a result, it is resistant to non-linear correlations and outliers. For large datasets, KNN can be computationally costly, particularly if there are numerous dimensions.

3.3.3. Support Vector Machines (SVM). A potent ML method for tasks like regression and classification is called support vector machines, or SVM. Finding a hyperplane in high-dimensional space that divides data points of different classes with the maximum margin is the fundamental notion behind SVM. The decision boundary is the name given to this hyperplane. Support vectors, or locations that are closest to the decision border, are essential for figuring out the decision function of the model. By transforming the data into a higher-dimensional space using kernel functions, SVM can handle linear and non-linear classification tasks. SVM is able to capture intricate feature correlations as a result. However, with large datasets, SVM can be computationally expensive, particularly when utilizing sophisticated kernel functions.

3.3.4. Gaussian Naïve Bayes (GNB). GNB is a probabilistic classification method based on the Bayes principle. It predicts that the features of the data are independent based on the class label. Probabilities are easier to calculate with the naïve Bayes assumption. GNB assumes that the features have a Gaussian (normal) distribution. Once the technique estimates the likelihood of each class given the input features, the data point is assigned to the class that has the highest probability. Text classification and spam filtering are two common applications for Gaussian Naive Bayes because of their efficiency and user-friendliness. It might not work as planned if the characteristics have a significant association.

3.3.5. Decision Tree (DT). Decision tree classification is used to create a decision-making tree model. Each node in the tree represents a test on a property, and each branch in the tree indicates a possible test outcome. The tree's leaves show the anticipated differences in class. Because DTs are easy to understand and illustrate, they are frequently utilized to help with the comprehension of the decision-making process. However, decision trees frequently overfit, especially in cases when the input contains noise or anomalies. One method that can assist in reducing this issue is pruning.

3.3.6. Random Forest. The powerful algorithm for ensemble learning To generate forecasts, Random Forest makes use of several decision trees. Every DT in the forest is trained using a distinct subset of the features and data to reduce the possibility of overfitting. This ensemble approach improves the model's accuracy and generalization performance. Random forests are helpful for many applications since they can handle numerical and category data. They are frequently used for tasks like outlier identification, regression, and classification in industries including natural language processing, finance, and healthcare. In addition to being more computationally expensive to train on large datasets, random forests might

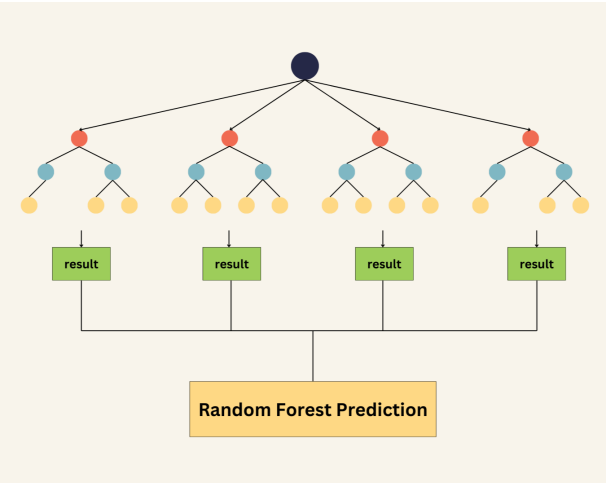

Figure 2: Random Forest Architecture

be harder to understand than individual DTs. The architecture of the random forest is shown simply in Fig.2

3.3.7. Gradient Boosting. Gradient boosting is a ML technique that focuses on fixing the mistakes from the previous stage as it develops a model piecemeal. It's a practical group technique that can successfully finish a range of assignments. Gradient boosting uses the residuals from the previous stage to train a weak learner, usually a decision tree. The final projection is then created by combining the forecasts from each phase. Gradient boosting is an iterative technique that enhances the performance of the model and makes it possible to find intricate patterns in the data. Applications including regression, ranking, and classification frequently use gradient boosting. However, computational costs associated with training can be high when the dataset is huge.

3.3.8. XGBoost. XGBoost's popularity and effectiveness stem from its robust implementation of gradient boosting principles, which combines the strengths of both boosting and gradient descent techniques. This approach enables XGBoost to achieve impressive accuracy and generalization capabilities across diverse datasets, while also offering mechanisms for fine-tuning model parameters to optimize performance. Its efficient handling of high-dimensional data and ability to capture intricate relationships between features make it particularly suitable for complex real-world problems in areas such as finance, healthcare, and marketing. Furthermore, XGBoost's interpretability features, such as feature importance ranking and visualization of decision trees, aid in understanding model predictions and gaining actionable insights from data. Continual updates and contributions from the community ensure that XGBoost remains a cornerstone in the field of machine learning, driving advancements that benefit both researchers and practitioners alike.

3.3.9. CatBoost. CatBoost, short for Categorical Boosting, is a gradient boosting algorithm specifically designed to handle categorical features efficiently. Developed by Yandex, CatBoost addresses common challenges in machine learning, such as dealing with categorical data, avoiding overfitting, and ensuring robust and accurate predictions. One of its standout features is the ability to handle categorical variables directly, without the need for extensive preprocessing or one-hot encoding, which simplifies the modelling process and reduces computational overhead. CatBoost uses a novel technique called "ordered boosting" to mitigate prediction shifts and achieve better generalisation. In addition to its advanced handling of categorical data, CatBoost includes several other features that enhance its performance and usability. It supports GPU training, which significantly speeds up the computation, especially for large datasets. The algorithm also incorporates built-in mechanisms for handling missing values and offers extensive hyperparameter tuning options to optimize model performance. CatBoost's effectiveness has been demonstrated in various competitions and practical applications, particularly in fields like finance, e-commerce, and

biology, where categorical data is prevalent. Its ease of use and high accuracy and robustness make CatBoost a powerful tool for data scientists and ML practitioners aiming to build reliable predictive models.

3.3.10. LightGBM. Gradient-based One-Side Sampling (GOSS) is a new approach that LightGBM, a gradient-boosting framework, employs to accelerate training. GOSS focuses on sampling data points with large gradients, which contribute more to the loss function. This reduces the number of data points that need to be considered during training, leading to faster convergence. LightGBM also uses exclusive feature binning, which creates histograms of feature values to calculate gradients efficiently. This further accelerates the training process. LightGBM is highly efficient and scalable, making it suitable for large datasets. It's often used for tasks like classification, regression, and ranking in various domains.

3.4. Proposed Model Description

3.4.1. Bagging (Random Forest). The first step in bagging is creating subsets of the dataset. The training data will be used to create random subsets of the data that will be used to train multiple models parallelly in bagging. The results obtained by the bagging are used to predict the final results. Using the algorithm Random Forest for the bagging technique is considered to be the best choice as the best accuracy was shown by this in many existing methods. Random Forest uses multiple decision trees as the base of the technique. For each subset of the dataset, a decision tree will be obtained and for the prediction results all the decision tree results will be averaged. There are many reasons to only choose to boost with Random Forest some of the main reasons are that reduces overfitting and also helps in improving accuracy and robustness. The process of averaging all the results in the process of bagging helps reduce the overfitting of the model, the decision trees if considered individually can show overfitting but combining them using random forest decreases the chances of overfitting. When compared to a single model, combining multiple models always gives the best accuracy. The usage of Random Forest with bagging also gives the advantage of the model being diverse as it can obtain feature randomness and helps in creating accurate models even with diverse datasets.

3.4.2. Boosting (XG Boost, Light GBM, CatBoost). Boosting is an ensemble learning technique, the most common use of boosting is using weak algorithms combining them and obtaining a strong model. Boosting builds the model from the base. The first step is to train a base model on the entire dataset, and then by correcting the errors another model is trained on the same dataset, this is an iterative process and works till the last model by correcting the combined error of all the previous models. The three algorithms that are used for the boosting technique are XG Boost, Light GBM, and CatBoost. XG Boost is popularly known to prevent overfitting, it uses gradient boosting with several optimization techniques, even when using large datasets, XG Boost can obtain the most accurate results rapidly and also can handle the missing data. Light GBM is used with a boosting technique as it can handle massive datasets with ease. It is specifically designed to give accurate results while maintaining high-performance rates. For results in the best accuracy, Light GBM grows the trees leaf-wise instead of level which uses gradient boosting. The dataset that is used for the study contains categorical values and the best algorithm to use for categorical data is CatBoost. CatBoost is known for its efficiency in handling categorical data without extensive preprocessing. This uses order boosting and reduces the overfitting chances. The training of CatBoost is also optimized in a way that gives the best accuracy rapidly. The main objective of boosting is to correct the errors of the previously trained models and this works best when used with multiple algorithms that are known for their best performance and accuracy rates. Including the boosting algorithm in the model helps in boosting accuracy, and flexibility, and handles massive data with much ease.

3.4.3. Stacking (Linear Regression, SVM). Stacking, also known as stacked generalization, is an ensemble learning strategy that enhances the performance of the primary model by combining several predictive models. Initially, all the selected models will be trained on the same dataset, and then all the results will be taken as inputs into a meta-model, The model is like the backbone of the stacking technique as it combines the results of all the models and gives the best result as a conclusion. The usage of Linear Regression and SVM as the base models for stacking gives a lot of advantages. The interpretability mechanism of linear regression helps in identifying the relations between the predicted variables and

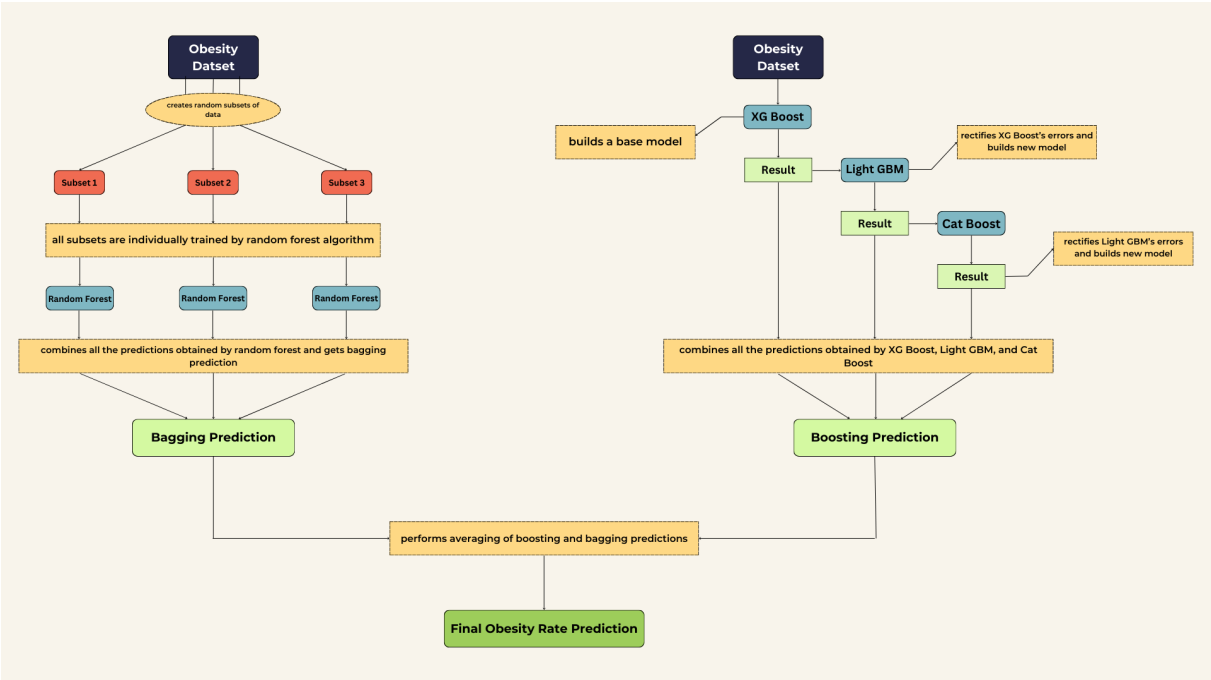

Figure 3: Hybridized Model Architecture

the target variables. The traditional linear regression is considered a strong base model as it creates a strong foundation for ensemble stacking. Support Vector Machine is a strong base model due to its precise accuracy in classification and prediction tasks. The usage of SVM is very useful while dealing with massive data as it can easily capture complex patterns in data, it is also robust to overfitting and outliers. Usually, SVMs give a different perspective of the data, and it is very useful in stacking techniques. The linear that might be missed by SVM will be handled by the Linear Regression. The main task when using these algorithms is efficiently optimizing the hyperparameters effectively as the final prediction will be based on the parameters. Implementing the stacking technique with different base models helps to improve the performance rate and diversification of the model.

3.4.4. Hybridization. Creating this hybrid model consists of three primary steps. The first step is to perform the bagging technique which is bootstrap aggregation, using the machine learning algorithm, Random Forest. Random Forest helps by increasing the accuracy and decreasing the overfitting. For the bagging technique, the training data will be sampled into different subsets and the random forest trains multiple decision trees on various subsets of data. The model can readily overcome any noise or outliers in the data by training on multiple subsets. It allows the trees not to be dependent on a small set of features because every decision tree in the forest is trained on a separate set of features. Another useful advantage of using random forest is it increases the accuracy rate as it combines the predictions of multiple decision trees, and the decisions of all trees are independent of each other, so parallelly multiple trees can be trained and this makes it efficient to use even for large datasets. The model trained with Random Forest is known to not use any underlying data to make assumptions. Hence using the random Forest algorithm with bagging technique is considered to be one of the best methods. The second step is performing the boosting technique on the same training data. Boosting is an ensemble technique and, in this model, the algorithms, XG Boost, Light GBM, and CatBoost are used to perform the boosting technique. It builds models iteratively by increasing accuracy in each step. Boosting technique is popularly known for learning from its previous model's mistakes. Most of the time boosting is used on weak algorithms to boost their performance. In the proposed model, the boosting is first performed on XG Boost, the model is trained on the same dataset that has been used for Random Forest, building a model from XG Boost is

the first step in boosting, it helps is regularization and gives promising results, the next step in boosting is to use the obtained results and build the next model using Light GBM, here, usage of boosting helps by finding out the mistakes in the previous model and improves them, the final step in boosting is to build a new model using CatBoost rectifying all the errors faced by the previous models.

The L1 and L2 regularization by the algorithms helps in reducing overfitting. Boosting these algorithms will also help in the early stopping mechanism, to stop training a model when performance is not improving or degrading. The final step of hybridization is merging the results obtained in step 1 and step 2, i.e., combining the results obtained by both, Random Forest with Bagging, and XG Boost, Light GBM, and CatBoost with Boosting. The whole architecture is explained in a diagrammatic way in Fig. 3. While merging the results there are various ways, The most popular methods are averaging, weighted averaging, and stacking, without further complexing the model, it is best to use the averaging method. This calculates the average of the predictions obtained by both models and gives the final predictions. Averaging can also reduce the variance in the prediction, which helps prevent the model from making extreme predictions. Also as more than three algorithms were used to build the base model, the usage of averaging method will help to mitigate the impact of individual model errors and can improve the overall accuracy rate of the combined model.

While the hybrid model improved prediction accuracy, it also introduced greater computational complexity due to the parallel and sequential training of multiple models. However, training times remained acceptable (less than 10 minutes) on a standard high-performance GPU, and prediction latency was within real-time application constraints. A trade-off analysis suggests the accuracy gain justifies the added computational load for clinical settings.

3.4.5. Hyperparameter Optimization. Each ensemble algorithm in the proposed model was optimized using grid search and cross-validation to enhance predictive performance. For the Random Forest model, the hyperparameters tuned included n_estimators, max_depth, and min_samples_split. In the case of XGBoost and LightGBM, the parameters learning_rate, n_estimators, and max_depth were optimized. For CatBoost, iterations, depth, and l2_leaf_reg were carefully adjusted. These hyperparameter optimizations played a crucial role in achieving the high accuracy reported by the model.

4. Result Analysis

Table 2: Performance Comparison of Various Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.86	0.86	0.87	0.87
K-Nearest Neighbors (KNN)	0.80	0.80	0.80	0.80
Support Vector Machine (SVM)	0.88	0.89	0.88	0.89
Gaussian Naive Bayes	0.57	0.57	0.57	0.52
Decision Tree	0.88	0.88	0.88	0.88
Random Forest	0.89	0.90	0.89	0.89
Gradient Boosting	0.27	0.38	0.29	0.26
XGBoost	0.91	0.91	0.91	0.91
CatBoost	0.90	0.91	0.90	0.91
LightGBM	0.91	0.91	0.91	0.91
Bagging with Random Forest	0.92	0.92	0.92	0.92
Boosting with XGBoost, CatBoost, LightGBM	0.94	0.94	0.94	0.94
Stacking (Linear Regression & SVM)	0.89	0.89	0.89	0.89
Proposed Model (Bagging + Boosting)	0.97	0.97	0.97	0.97

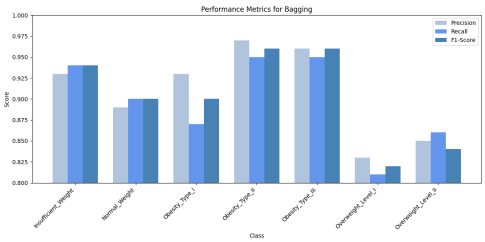

Figure 4: Classification graph of Bagging Model

When creating a model, the most crucial stage is to assess it and compare the outcomes with those of other models that are already in use. The model was evaluated using the four primary performance indicators: F1-Score, Accuracy, Precision, and Recall. Recall is the proportion of real positives that are true positives; precision is the proportion of positive forecasts that are true positives; and accuracy is the proportion of correctly expected occurrences. These metrics are used to evaluate the model. The f1-score strikes a balance between recall and precision by using their harmonic mean. In this study, the developed model was compared to 10 other existing models and 3 new models, the proposed hybrid model showed a great performance among all the other models while using the same dataset. Among these, the least accuracy was achieved by algorithms like Gradient Boosting and Gaussian Naive Bayes, the highest accuracy achieved by the existing algorithms is 91% from XG Boost and Light GBM. In the new models, the highest accuracy achieved is 94% and the Proposed hybrid model obtained an accuracy of 97%. Even when comparing the model using other performance metrics such as Precision, Recall, and F1-score, the proposed model still stands as the best one. All the values are shown in Table. 2

Tables 3 4 5 6 and Fig. 4 5 6 7 present the classification performance metrics across four different methods: Bagging, Boosting, Stacking, and the Proposed Hybrid Model. The values captured in these tables allow for a comprehensive comparison of each method's effectiveness, providing insights into their respective strengths and potential areas for improvement. Upon analyzing the metrics, it becomes evident that the Proposed Hybrid Model consistently outperforms the other methods across all evaluated criteria. This superior performance underscores the enhancements and innovations incorporated into the hybrid approach, setting it apart from traditional ensemble techniques like Bagging, Boosting, and Stacking.

Table 3: Classification performance using Bagging

Class	**Precision**	**Recall**	**F1-Score**
Class_1	0.93	0.94	0.94
Class_2	0.89	0.90	0.90
Class_3	0.93	0.87	0.90
Class_4	0.97	0.95	0.96
Class_5	0.96	0.95	0.96
Class_6	0.83	0.81	0.82
Class_7	0.85	0.86	0.84

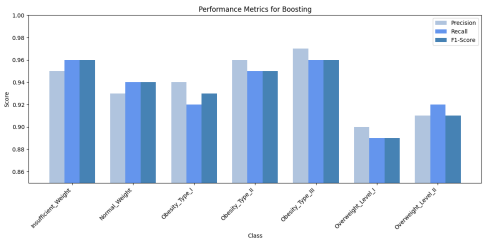

Figure 5: Classification graph of Boosting Model

Table 4: Classification performance using Boosting

Class	Precision	Recall	F1-Score
Class_1	0.95	0.96	0.96
Class_2	0.93	0.94	0.94
Class_3	0.94	0.92	0.93
Class_4	0.96	0.95	0.95
Class_5	0.97	0.96	0.96
Class_6	0.90	0.89	0.89
Class_7	0.91	0.92	0.91

Table 5: Classification performance using Stacking

Class	Precision	Recall	F1-Score
Class_1	0.89	0.90	0.89
Class_2	0.87	0.88	0.87
Class_3	0.89	0.85	0.87
Class_4	0.91	0.89	0.90
Class_5	0.92	0.91	0.91
Class_6	0.82	0.80	0.81
Class_7	0.83	0.84	0.83

Table 6: Classification performance using the Hybridized Model

Class	Precision	Recall	F1-Score
Class_1	0.98	0.98	0.98
Class_2	0.96	0.97	0.97
Class_3	0.97	0.95	0.96
Class_4	0.98	0.97	0.97
Class_5	0.99	0.98	0.98
Class_6	0.94	0.93	0.94
Class_7	0.95	0.95	0.95

Fig. 7 presents a bar graph that illustrates the classification performance across all classes in the proposed model, evaluated using the key performance metrics: Precision, Recall, and F1-Score. The graph indicates that the model achieves consistently high scores across all metrics for each class. The uniformity in the classification results across these diverse categories suggests that the model is both robust and reliable in distinguishing between different weight-related conditions. The nearly identical performance across all categories is a significant factor contributing to the model's overall high accuracy rate of 0.97. Such uniformity not only underscores the model's precision but also highlights its potential for reliable application in real-world scenarios where accurate classification is critical. The proposed model's ability to maintain high performance across diverse classes justifies its potential as a powerful tool for weight classification tasks, providing consistent and trustworthy results. The values are given in Table. 6

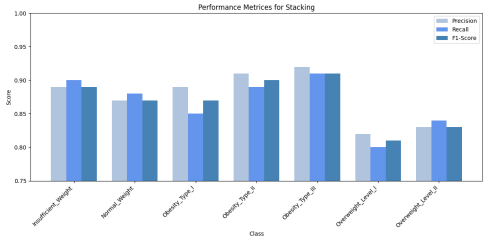

Figure 6: Classification graph of Stacking Model

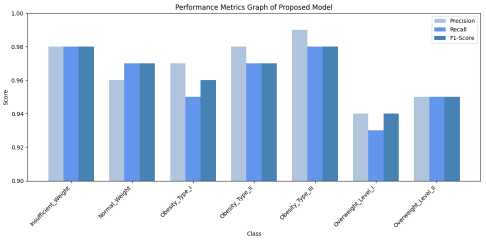

Figure 7: Classification graph of Hybridized Model

Fig. 8 shows a correlation heatmap, which visualizes the relationships between different features in the dataset. The colour intensity in the image is used to indicate the strength and direction of the correlation and the correlation coefficient between two features is represented by each cell in the heatmap. A correlation coefficient close to 1 (lighter colours) suggests a strong positive correlation, while a coefficient close to -1 (darker colours) indicates a strong negative correlation. The diagonal line of white squares represents the perfect correlation of each feature with itself (correlation = 1). In this specific heatmap, there are no extremely strong correlations between most features, as indicated by the predominance of reddish colours, which correspond to weaker correlations. However, certain clusters or pairs of features show slightly higher correlations, which could indicate relationships worth exploring further in analysis. For instance, categories related to gender, family history, and certain obesity classifications might display patterns of correlation with other lifestyle or physical attributes. This heatmap is useful for identifying which features are most interrelated, helping to guide feature selection and understanding potential multicollinearity in predictive modelling.

Fig. 9 presents histograms with density plots for eight features: Age, Height, Weight, FCVC (vegetable consumption frequency), NCP (number of daily meals), CH2O (daily water intake), FAF (physical activity frequency), and TUE (time using technology). The distributions reveal key insights: Age is right-skewed, indicating a younger population, while Height and Weight follow a normal distribution. FCVC shows a high frequency of vegetable consumption, and NCP spikes at three meals per day, reflecting common eating habits. CH2O and FAF have multimodal distributions, suggesting varied water intake and activity levels. TUE shows most participants with low to moderate technology use, offering a detailed snapshot of lifestyle and physical characteristics within the dataset.

Fig. 10 is a pair plot showing the relationships between multiple numerical features in a dataset, categorized by obesity levels. Each subplot within the grid represents a scatter plot or distribution plot, illustrating the interaction between two variables. The diagonal plots show how each variable is distributed, while the off-diagonal plots show how pairs of variables are related to each other. The data points are color-coded according to different obesity levels, ranging from 0 to 6, allowing for visual comparison of how these variables differ across obesity levels. This visualization helps identify correlations and patterns between the features and obesity levels.

Fig. 11 displays histograms of log-transformed distributions for several variables, including Age, Height, Weight, and others. Each subplot represents the distribution of one variable after applying a logarithmic transformation, which is typically done to normalise skewed data and make patterns more discernible. The histograms show the frequency (count) of observations across different ranges of the log-transformed variable, with a smooth density curve overlaying each histogram to illustrate the distribution's shape. This transformation is useful for understanding the underlying patterns and spread of data, especially when the original data might be heavily skewed.

The model's prediction capabilities could be used in preventive healthcare systems to identify individuals at high obesity risk early, enabling timely lifestyle interventions. Feature importance analysis from the ensemble methods highlighted that dietary habits (e.g., FCVC, physical activity (FAF), and meal frequency (NCP)) were the most influential predictors, aligning with clinical understanding of obesity risk factors.

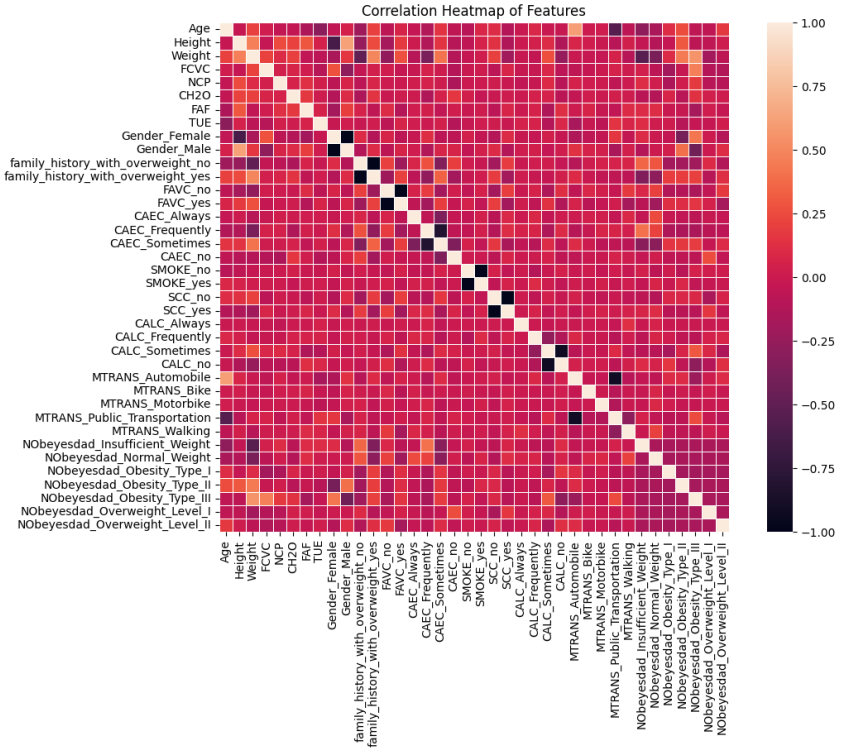

Figure 8: Correlation Heatmap

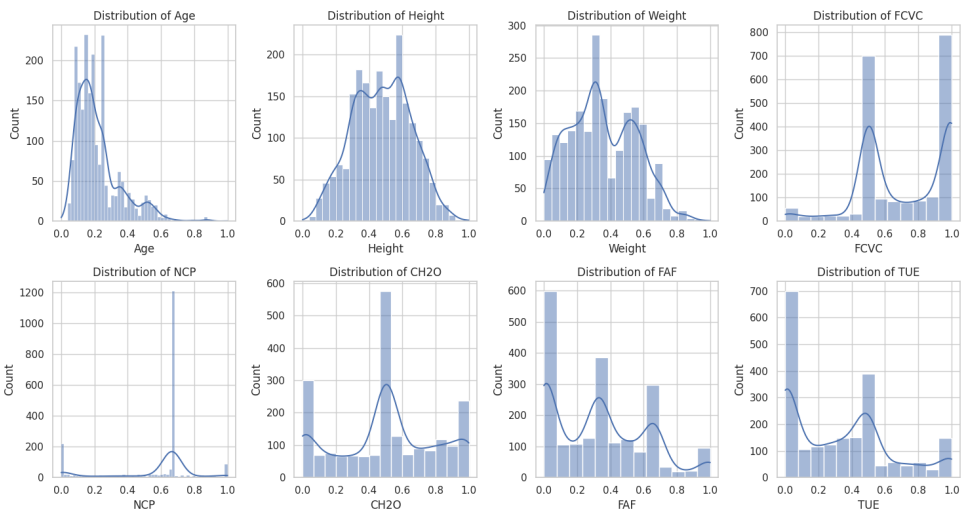

Figure 9: Distribution plots for Various Features in the Dataset

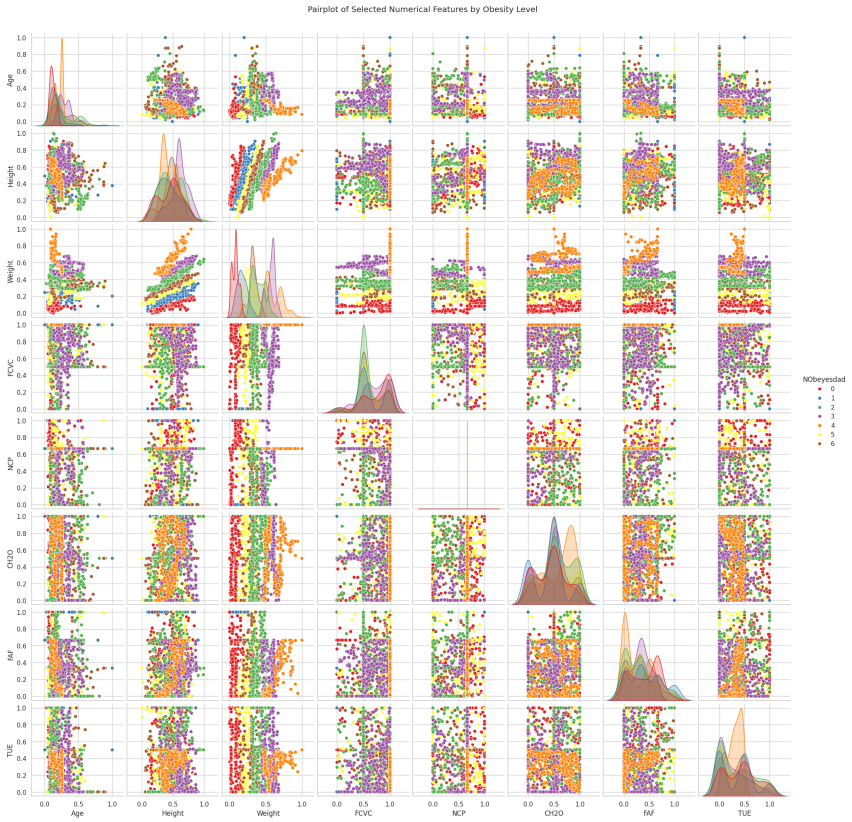

Figure 10: Pairplot of Selected Numerical Features by Obesity Level

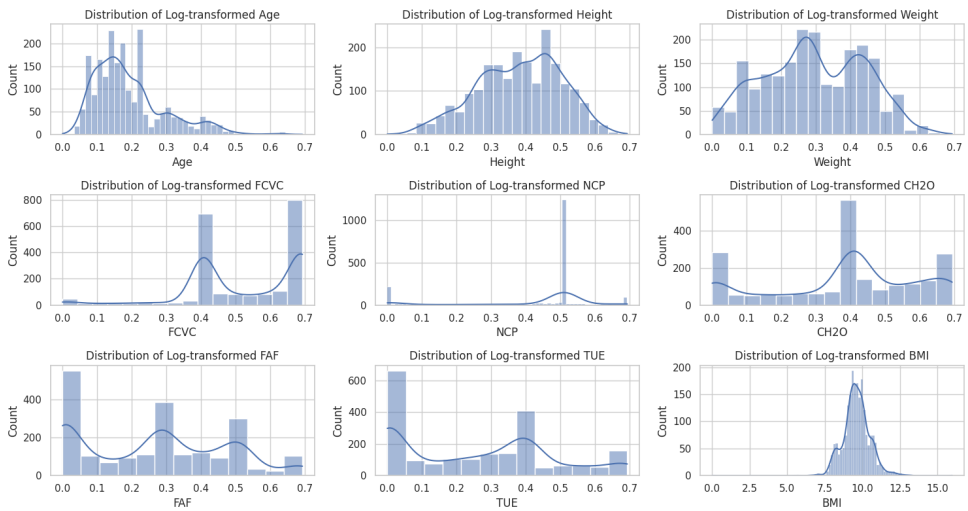

Figure 11: Distribution Plots for Various Features in the Dataset Using Log Transformation

5. Conclusion

This study presents a robust hybrid ensemble model integrating Bagging and Boosting techniques, achieving 97.85% accuracy in obesity risk classification. Through methodical preprocessing, hyperparameter tuning, and a thoughtful ensemble design, the model surpasses traditional baselines. Its reliability across all obesity classes and strong interpretability via feature importance metrics establish it as a practical tool for healthcare analytics. Future work could incorporate genetic and environmental data to enhance prediction depth and real-world applicability.

In future, this study can be extended more by diving deep data-wise and creating a massive dataset that consists of various features like the environmental factors and obese genes in the ancestors. Another way to extend the study is by identifying the limitations of the existing work and developing a model that can rectify them.

Acknowledgments

Authors are thankful to the editor & learned referee for their inspiring and fruitful suggestions. The authors are also thankful to the Vellore Institute of Technology, Andhra Pradesh, Amaravati, for providing the necessary infrastructure for the completion of the current research work.

References

1. Harold Edward Bays, Shagun Bindlish, and Tiffany Lowe Clayton. Obesity, diabetes mellitus, and cardiometabolic risk: an obesity medicine association (oma) clinical practice statement (cps) 2023. *Obesity Pillars*, 5:100056, 2023.

2. Bryan Chong, Jayanth Jayabaskaran, Gwyneth Kong, Yiong Huak Chan, Yip Han Chin, Rachel Goh, Shankar Kannan, Cheng Han Ng, Shaun Loong, Martin Tze Wah Kueh, et al. Trends and predictions of malnutrition and obesity in 204 countries and territories: an analysis of the global burden of disease study 2019. *EClinicalMedicine*, 57, 2023.

3. Yan-Hui Dong, Li Chen, Jie-Yu Liu, Tao Ma, Yi Zhang, Man-Man Chen, Pan-Liang Zhong, Di Shi, Pei-Jin Hu, Jing Li, et al. Epidemiology and prediction of overweight and obesity among children and adolescents aged 7-18 years in china from 1985 to 2019. *Zhonghua yu Fang yi xue za zhi [Chinese Journal of Preventive Medicine]*, 57:11–19, 2023.

4. Arielle Elmaleh-Sachs, Jessica L Schwartz, Carolyn T Bramante, Jacinda M Nicklas, Kimberly A Gudzune, and Melanie Jay. Obesity management in adults: a review. *Jama*, 330(20):2000–2015, 2023.

5. Faria Ferdowsy, Kazi Samsul Alam Rahi, Md Ismail Jabiullah, and Md Tarek Habib. A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, 2:100053, 2021.

6. Hao Gou, Huiling Song, Zhiqing Tian, and Yan Liu. Prediction models for children/adolescents with obesity/overweight: A systematic review and meta-analysis. *Preventive Medicine*, page 107823, 2023.

7. Junhwi Jeon, Sunmi Lee, and Chunyoung Oh. Age-specific risk factors for the prediction of obesity using a machine learning approach. *Frontiers in Public Health*, 10:998782, 2023.

8. Ji-Hoon Jeong, In-Gyu Lee, Sung-Kyung Kim, Tae-Eui Kam, Seong-Whan Lee, and Euijong Lee. Deephealthnet: Adolescent obesity prediction system based on a deep learning framework. *IEEE Journal of Biomedical and Health Informatics*, 2024.

9. Markus Jokela and Michael Laakasuo. Obesity as a causal risk factor for depression: systematic review and meta-analysis of mendelian randomization studies and implications for population mental health. *Journal of psychiatric research*, 163:86–92, 2023.

10. Holly Lofton, Jamy D Ard, Rameck R Hunt, and Michael G Knight. Obesity among african american people in the united states: A review. *Obesity*, 31(2):306–315, 2023.

11. Krishna Modi, Ishbir Singh, and Yogesh Kumar. A comprehensive analysis of artificial intelligence techniques for the prediction and prognosis of lifestyle diseases. *Archives of Computational Methods in Engineering*, 30(8):4733–4756, 2023.

12. Pritom Kumar Mondal, Kamrul H Foysal, Bryan A Norman, and Lisaann S Gittner. Predicting childhood obesity based on single and multiple well-child visit data using machine learning classifiers. *Sensors*, 23(2):759, 2023.

13. Casimiro Aday Curbelo Montañez, Paul Fergus, Abir Hussain, and Al-Jumeily. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2743–2750. IEEE, 2017.

14. Balbir Singh and Hissam Tawfik. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, pages 523–535. Springer, 2020.

15. Dahlak Daniel Solomon, Shakir Khan, Sonia Garg, Gaurav Gupta, Abrar Almjally, and Alabduallah. Hybrid majority voting: Prediction and classification model for obesity. *Diagnostics*, 13(15):2610, 2023.

16. Fatma Hilal Yagin, Mehmet Gülü, Yasin Gormez, Arkaitz Castañeda-Babarro, Cemil Colak, Gianpiero Greco, Francesco Fischetti, and Stefania Cataldi. Estimation of obesity levels with a trained neural network approach optimized by the bayesian technique. *Applied Sciences*, 13(6):3875, 2023.

Sai Charan Medaramatla,
Department of School of Computer Science and Engineering,
VIT-AP UNIVERSITY,
Amaravati-522237, India.
E-mail address: medaramatla.saicharan@gmail.com

and

Shekshavali Pattan,
Department of School of Computer Science and Engineering,
VIT-AP UNIVERSITY,
Amaravati-522237, India.
E-mail address: shekshavalipattan14@gmail.com

and

Sai Harshavardhan Srungavarapu,
Department of School of Computer Science and Engineering,
VIT-AP UNIVERSITY,
Amaravati-522237, India.
E-mail address: saiharshavardhanshv2003@gmail.com

and

Voddelli SriLakshmi,
Department of School of Computer Science and Engineering,
VIT-AP UNIVERSITY,
Amaravati-522237, India.
E-mail address: srilakshmi.v@vitap.ac.in