BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

Implementation of the LightGBM–CatBoost Ensemble Method for Obesity Risk Classification in Productive Age

Kecitaan Harefa*, Joko Priambodo

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1,*]dosen00842@unpam.ac.id, [2]dosen00276@unpam.ac.id
Correspondence Author Email: dosen00842@unpam.ac.id

Abstract−Obesity is a health problem that continues to increase among individuals of productive age and has the potential to reduce quality of life and work productivity. One of the main challenges in obesity risk assessment is the limitation of conventional methods in accurately identifying obesity risk when dealing with complex, multidimensional data that include both numerical and categorical variables. Therefore, an artificial intelligence–based approach is required to provide a more accurate and stable obesity risk classification. This study aims to implement and evaluate a LightGBM–CatBoost ensemble method for obesity risk classification with a focus on the productive age population. The dataset used in this study was obtained from the Kaggle platform and consisted of 2,111 individual records containing physical attributes, eating habits, physical activity, and lifestyle factors. Although the dataset is synthetic and balanced, the included attributes and age-related variables are representative of individuals within the productive age range, making it suitable for modeling obesity risk in this demographic context. The research stages include data preprocessing, separate training of the LightGBM and CatBoost models, model integration using a probability averaging ensemble technique, and performance evaluation using accuracy, precision, recall, and F1-score metrics. The results indicate that both LightGBM and CatBoost achieved accuracy levels above 95%, while the ensemble model demonstrated superior performance with an accuracy of 96.69% and more balanced evaluation metrics across all obesity risk classes. These findings confirm that the ensemble approach improves classification stability and accuracy compared to single models. Therefore, the LightGBM–CatBoost ensemble method is effective for obesity risk classification and has the potential to be further developed as a decision support system in the health sector.

Keywords: Obesity; Ensemble Learning; LightGBM; CatBoost; Risk Classification

1. INTRODUCTION

Obesity is one of the global health problems whose prevalence continues to increase from year to year and is a serious concern in various countries, both developed and developing countries [1]. The condition of obesity is characterized by excessive accumulation of body fat and has the potential to cause various chronic diseases such as type 2 diabetes mellitus, cardiovascular disease, hypertension, and other metabolic disorders [2]. Obesity measurement is generally carried out using Body Mass Index (BMI), but in practice, obesity is also influenced by many other factors, such as body fat distribution, diet, physical activity level, lifestyle, and genetic factors [3]. In the past decade, the increase in obesity cases has been seen significantly in the productive age group, which should be in optimal health conditions to support work productivity and economic development [4].

The increase in obesity in the productive age is closely related to changes in modern lifestyles, such as increased consumption of high-calorie fast food, low physical activity due to sedentary work patterns, and irregular sleep habits [5]. In addition, behavioral factors such as exercise frequency, water consumption, and smoking habits also contribute to obesity risk [6]. Recent studies show that genetic factors and family history of obesity also affect a person's tendency to be obese [7]. In this context, obesity-related data typically consist of a combination of anthropometric attributes (such as height, weight, and BMI-related indicators) and behavioral or lifestyle factors, including dietary habits, physical activity patterns, and daily routines. The complexity of these factors causes obesity risk analysis to become a multidimensional problem that is difficult to solve using conventional approaches [8]. Therefore, an analysis method is needed that is able to process data with many variables, both numerical and categorical, and is able to capture nonlinear relationships between causative factors [9].

Classic statistical approaches, such as linear regression or logistic regression, are still widely used in health research to predict obesity risk [10]. However, the method has limitations in handling data with high dimensions and complex variable relationships. In addition, conventional methods tend to assume a linear relationship between variables, whereas health and lifestyle data are often nonlinear and influenced by interactions between features. This limitation can lead to low prediction accuracy and less stability when faced with heterogeneous data. Therefore, the use of artificial intelligence techniques, especially machine learning, is an alternative that is increasingly used to improve the quality of obesity risk prediction [11].

Various studies have applied machine learning algorithms for the classification and prediction of obesity, such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, and XGBoost [12]. The results show that decision tree-based algorithms and boosting generally produce better performance than linear methods. Random Forest, for example, has been proven to be able to provide high accuracy to health and lifestyle data due to its ability to reduce overfitting through an internal ensemble mechanism. However, some studies have also reported that Random Forest still has limitations in handling complex categorical features and requires additional preprocessing. Other research began to adopt modern boosting algorithms such as LightGBM and CatBoost that offered better computing efficiency and performance, but most still used a single model without combining the advantages of multiple algorithms at once [13].

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

The Light Gradient Boosting Machine (LightGBM) is a boosting algorithm designed to improve the efficiency and speed of model training, especially on large datasets [14]. LightGBM uses histogram-based decision tree techniques and leaf-wise growth strategies that are able to produce high accuracy with relatively low computation time [15]. On the other hand, CatBoost was developed to address the problem of categorical data that often arises in real-world data, including health and lifestyle data [16]. CatBoost has an effective internal encoding mechanism and is able to reduce the risk of overfitting through the ordered boosting technique [17]. The combination of these two algorithms is considered appropriate to handle the characteristics of obesity data consisting of a mixture of numerical and categorical variables [18].

The ensemble approach aims to combine multiple predictive models to produce better performance than a single model. By combining LightGBM and CatBoost, it is hoped that a more stable, accurate, and robust model can be obtained for data variations [19]. LightGBM excels at handling numerical features and computational efficiency, while CatBoost has advantages in categorical feature processing and overfitting prevention [20]. Through the ensemble technique based on probability aggregation, the weaknesses of each model can be minimized, and its advantages can be maximized. This approach is in line with the latest research trends that show that ensemble learning is able to improve classification performance on various complex problems, including in the health sector.

Based on the background and previous research studies, it can be concluded that the classification of obesity risk at productive age requires an approach that is able to handle multidimensional and complex data effectively [21]. Therefore, this study aims to implement the LightGBM–CatBoost ensemble method in the classification of obesity risk as well as evaluate its performance compared to a single model. It is hoped that the results of this research can make a scientific contribution to the development of machine learning methods for obesity prediction, as well as become the basis for the development of decision support systems that can help efforts to prevent and control obesity in the community.

2. RESEARCH METHODOLOGY

This research method outlines the main steps for developing an obesity risk classification model using the LightGBM–CatBoost ensemble approach. As shown in Figure 1, the procedure includes data collection, data preprocessing, training the LightGBM and CatBoost models, combining both models using probability averaging, and evaluating performance using classification metrics [22].

Figure 1. Research Methods

The following is an explanation of the steps (Figure 1) of the method above:

a. Dataset Collection
   The process of collecting data is used as research material. At this stage, obesity datasets are collected from reliable sources that contain information related to individual characteristics, such as demographic data, eating habits, physical activity, and other lifestyle factors related to obesity risk. The dataset obtained was then selected to ensure the completeness of the attributes and the suitability with the research objectives, namely the classification of obesity risk at productive age.

b. Data Preprocessing
   The data preprocessing stage aims to prepare the dataset so that it can be effectively used in the modeling process. At this stage, the dataset was first examined for missing values, and the results showed that no missing data were present, so no imputation was required. Categorical variables, such as gender, eating habits, family history of obesity, and transportation mode, were transformed into numerical representations using Label Encoding to ensure compatibility with the LightGBM and CatBoost algorithms. Because both LightGBM and CatBoost are tree-based models that are insensitive to feature scaling, data normalization, or standardization was not applied. The dataset was then divided into training and testing sets using an 80:20 split to allow objective evaluation of model performance. For model configuration, the main hyperparameters, including the number of trees, learning rate, and tree depth, were determined through preliminary experimentation based on commonly used values in related studies, without employing automated

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

hyperparameter optimization techniques such as Grid Search or Random Search. This preprocessing strategy ensures data consistency while maintaining computational efficiency and model reliability.

c. LightGBM Model Training
The Light Gradient Boosting Machine algorithm is used to train classification models based on training data. LightGBM works by building a decision tree in stages using a gradient boosting approach, so that it is able to capture nonlinear patterns in the data efficiently. The resulting model is expected to be able to provide accurate early predictions of obesity risk.

d. CatBoost Model Training
The same data is used to train the model using the CatBoost algorithm. CatBoost has the advantage of handling categorical variables and reducing the risk of overfitting through an ordered boosting mechanism. By training the CatBoost model separately, other predictions that are complementary to the results of LightGBM are obtained.

e. Ensemble Model (Averaging)
The probability predictions from the LightGBM and CatBoost models are averaged (probability averaging) to produce the final prediction. This ensemble approach aims to take advantage of the advantages of each model so that it can improve the accuracy and stability of the classification results compared to the use of a single model.

f. Results and Conclusion
In the model evaluation, the performance of the system was tested using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The results of the evaluation were then analyzed to assess the effectiveness of the proposed ensemble method. Based on the test results, conclusions were drawn about the model's ability to classify obesity risk at a productive age and its potential application as a decision support system in the health sector.

3. RESULT AND DISCUSSION

This section explains in detail the stages of applying the proposed approach to solve the research problem, namely, developing a robust and accurate obesity risk classification model for productive-age individuals using multidimensional lifestyle and physical attribute data. In addition to reporting the final performance metrics, this section discusses how each stage of the pipeline contributes to the overall objective, why the selected algorithms are appropriate for the characteristics of the dataset, and how the evaluation results support the conclusion that the LightGBM–CatBoost ensemble approach provides a stable and reliable solution for multi-class obesity risk classification.

3.1 Overview of the Application Stages to Solve the Research Problem

The research problem addressed in this study involves classifying obesity risk into seven categories, which is a more challenging task than binary classification because class boundaries are often close, and some classes share similar anthropometric and behavioral patterns. The proposed solution applies a structured machine learning pipeline consisting of: (1) dataset acquisition and verification, (2) data preprocessing and feature transformation, (3) training individual models (LightGBM and CatBoost), (4) integrating both models into an ensemble using probability averaging, and (5) evaluating performance with class-sensitive metrics and error analysis tools. Each stage is designed to ensure that the model handles the complexity of the features and produces stable predictions across all classes.

First, the dataset was obtained from Kaggle and contains 2,111 individual records. Each record represents a combination of physical attributes and lifestyle-related variables, including dietary behaviors and physical activity indicators. This mixture of feature types (numerical and categorical) is a key source of complexity, as obesity risk is influenced not by one variable alone but by a combination of interacting factors. Second, preprocessing was conducted to ensure that the data could be processed by boosting-based tree models. Third, two strong gradient boosting algorithms were trained separately to provide complementary predictive “views” of the same data. Finally, the two models were merged through probability averaging to reduce variance and mitigate model-specific errors, leading to improved robustness and generalization.

3.2 Data Preprocessing: Preparing Multidimensional Data for Modeling

At the preprocessing stage, the dataset (2,111 samples) was analyzed to ensure quality, completeness, and suitability for predictive modeling. The first task was checking missing values across all attributes. The results confirmed that the dataset contained no missing values, meaning that no imputation procedures were needed. This is important because missing value handling can introduce bias or variance depending on the chosen imputation strategy; therefore, having a complete dataset supports a more stable training process and reduces uncertainty introduced by preprocessing choices.

The second preprocessing task involved transforming categorical variables into numerical representations. Obesity-related data commonly contains categorical variables such as gender, eating habit categories, family history indicators, and transportation mode. These variables cannot be processed directly by many machine learning algorithms unless they are encoded. In this study, categorical features were converted using Label Encoding, transforming each category into an integer representation. Label Encoding was applied consistently for all categorical columns to maintain uniform preprocessing across both LightGBM and CatBoost, enabling fair comparison and smooth integration in the ensemble stage.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

Notably, normalization or standardization was not applied because LightGBM and CatBoost are tree-based models. Tree-based learners are generally insensitive to feature scaling because they make decisions based on split thresholds and ordering rather than distance-based measures. Avoiding unnecessary scaling reduces preprocessing complexity, maintains interpretability of feature values, and ensures the pipeline remains efficient.

After encoding, the dataset was split into training and testing subsets using an 80:20 ratio, producing 423 test instances. The use of a train–test split is essential for objective evaluation because it allows performance measurement on unseen data, reflecting the model's generalization ability. Moreover, because the research problem includes seven classes, maintaining representative distributions during splitting is important to avoid evaluation bias. The resulting test set size is sufficient to evaluate multi-class metrics and to observe patterns of misclassification in the confusion matrix. Overall, this preprocessing stage produced structured, ready-to-use data and established the foundation for reliable model training and evaluation.

3.3 LightGBM Model Training: Application of the First Boosting Algorithm

Following preprocessing, the first model trained was LightGBM (Light Gradient Boosting Machine). LightGBM is a gradient boosting framework that sequentially builds decision trees, where each new tree attempts to correct the errors made by the previous trees. This mechanism is especially suitable for obesity risk prediction because the relationships between features and obesity categories are often nonlinear and influenced by interactions among multiple factors.

A key characteristic of LightGBM is its leaf-wise tree growth strategy, which grows trees by splitting the leaf that yields the largest reduction in loss. This strategy allows LightGBM to capture complex patterns efficiently, especially when the dataset contains interacting variables. In obesity risk classification, such interactions may include combinations of eating frequency, physical activity level, and demographic indicators that jointly influence class membership. As such, LightGBM serves as a strong candidate for modeling the multidimensional nature of obesity risk.

The evaluation results show that the LightGBM model achieved an accuracy of 96.45%, indicating excellent overall classification performance. Additionally, the precision, recall, and F1-score values per class ranged approximately from 0.91 to 1.00, reflecting a strong balance between identifying correct classes and minimizing misclassification. The macro average and weighted average values of 0.96 suggest that LightGBM performs consistently across classes despite potential differences in class frequencies. With 423 test samples, these results indicate that LightGBM generalizes well and can reliably classify obesity risk levels based on the given lifestyle and physical attributes.

From a methodological perspective, these results demonstrate that LightGBM is effective in addressing the research problem. However, while LightGBM performs strongly, the study also considers the possibility that a single model may still produce class-specific confusion in a multi-class setting, especially when classes are adjacent and share similar characteristics. Therefore, CatBoost was trained as a second model to provide a complementary learning mechanism and improve robustness through ensemble integration.

The detailed classification report of the LightGBM model is presented in Figure 2. As shown in Figure 2, the model achieves an overall accuracy of 0.96 and demonstrates strong class-wise precision, recall, and F1-score values across all seven obesity risk classes, supporting its reliability on the 423 test samples.

```
=== LIGHTGBM ===
Accuracy: 0.9645390070921985
              precision    recall  f1-score   support

           0       0.96      0.94      0.95        54
           1       0.87      0.95      0.91        58
           2       1.00      0.97      0.99        70
           3       0.98      1.00      0.99        60
           4       1.00      0.98      0.99        65
           5       0.95      0.91      0.93        58
           6       0.98      0.98      0.98        58

    accuracy                           0.96       423
   macro avg       0.96      0.96      0.96       423
weighted avg       0.97      0.96      0.96       423
```

Figure 2. LightGBM Model Training

3.4 CatBoost Model Training: Application of the Second Boosting Algorithm

The second model trained in this study was CatBoost (Categorical Boosting). CatBoost is also a gradient boosting algorithm, but is particularly known for handling categorical data effectively and preventing overfitting through specialized mechanisms. Two prominent characteristics of CatBoost are the use of symmetric trees and ordered boosting. Symmetric trees enforce consistent splits at the same depth, promoting stability, while ordered boosting reduces prediction bias and helps avoid target leakage during the boosting process.

CatBoost is highly relevant in this study because obesity datasets include a mixture of numerical features (e.g., physical attributes) and categorical features (e.g., lifestyle behaviors represented in categories). Even when categorical variables are encoded numerically, CatBoost's learning mechanism and structural regularization can reduce overfitting and provide stable performance across classes.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

The CatBoost model achieved an accuracy of 95.74%, which also reflects excellent classification capability. The precision, recall, and F1-score values for most classes were between 0.90 and 1.00, indicating that CatBoost recognizes obesity risk patterns consistently. Macro and weighted averages reached 0.96, confirming balanced performance across classes. These outcomes show that CatBoost is effective and reliable for obesity risk classification, particularly in datasets where categorical behaviors and routines play an important role.

Although CatBoost's accuracy is slightly lower than LightGBM in this experiment, it provides a different perspective on the data due to its symmetric tree structure and ordered boosting mechanism. This difference is valuable for ensemble learning because ensemble performance depends not only on individual model strength but also on diversity in learning patterns. In other words, two models that learn differently can complement each other by correcting one another's errors, thereby improving the final prediction stability.

As shown in Figure 3, CatBoost achieves an accuracy of 95.74% and maintains generally high precision, recall, and F1-score values across the seven obesity risk classes on the 423 test samples.

```
=== CATBOOST ===
Accuracy: 0.9574468085106383
              precision    recall  f1-score   support

           0       0.98      0.93      0.95        54
           1       0.86      0.95      0.90        58
           2       0.97      0.97      0.97        70
           3       1.00      0.98      0.99        60
           4       1.00      0.98      0.99        65
           5       0.94      0.88      0.91        58
           6       0.95      1.00      0.97        58

    accuracy                           0.96       423
   macro avg       0.96      0.96      0.96       423
weighted avg       0.96      0.96      0.96       423
```

Figure 3. CatBoost Model Training

3.5 Ensemble Model (Averaging): Integrating Complementary Models to Improve Robustness

After training LightGBM and CatBoost separately, the study applied the main proposed method: the LightGBM–CatBoost ensemble using probability averaging. This ensemble stage is crucial for solving the research problem because multi-class obesity classification requires not only accuracy but also stability and reliability across all seven categories. In probability averaging, both trained models generate probability estimates for each class, and these probabilities are averaged to produce a final prediction. The class with the highest average probability is selected as the predicted label.

The ensemble model achieved the highest accuracy of 96.69%, demonstrating an improvement over both individual models. Precision, recall, and F1-score values for most classes were in the range of 0.93 to 1.00, indicating strong and consistent class-level performance. Macro and weighted average values increased to 0.97, supporting the conclusion that the ensemble model is not only accurate but also stable across different obesity categories.

The improved performance can be explained by the complementary learning characteristics of the two models. LightGBM's leaf-wise growth strategy captures complex interactions among numerical features and can form highly discriminative decision boundaries. CatBoost's symmetric trees and ordered boosting contribute to more stable pattern learning, particularly in the presence of categorical variables and potential bias risks. Because these algorithms learn differently, their errors are not perfectly correlated. By averaging their probability outputs, the ensemble reduces individual model variance and mitigates errors that might occur if only one model were used. This mechanism strengthens prediction reliability and supports the objective of producing a robust obesity risk classification model.

As shown in Figure 4, the LightGBM–CatBoost ensemble achieves an accuracy of 96.69% with consistently high precision, recall, and F1-score values across all seven classes on the 423 test samples.

```
=== ENSEMBLE (LGBM + CATBOOST) ===
Accuracy: 0.966903073286052
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        54
           1       0.88      0.98      0.93        58
           2       0.99      0.97      0.98        70
           3       0.98      0.98      0.98        60
           4       1.00      0.98      0.99        65
           5       0.95      0.91      0.93        58
           6       0.98      0.98      0.98        58

    accuracy                           0.97       423
   macro avg       0.97      0.97      0.97       423
weighted avg       0.97      0.97      0.97       423
```

Figure 4. Ensemble Model (Averaging)

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

3.6 Multi-Class Evaluation: Why Accuracy Alone Is Not Sufficient (7 Classes)

Although the ensemble model achieved high accuracy, it is important to interpret the results appropriately. In a seven-class classification problem, relying solely on accuracy can be misleading because accuracy does not show which classes are being confused. A model could achieve high accuracy while consistently misclassifying specific class pairs—especially classes that are adjacent and share similar feature profiles. Therefore, class-wise evaluation and error analysis are necessary to fully explain the performance of the model.

In this experiment, the ensemble accuracy of 96.69% corresponds to an error rate of 3.31%, which equals approximately 14 misclassified samples out of 423 test instances. To identify where these errors occur and how serious they are, the study used a Confusion Matrix, presented as a heatmap. This visualization provides detailed information about the distribution of correct predictions (diagonal cells) and misclassifications (off-diagonal cells).

The confusion matrix heatmap is particularly valuable because it highlights whether the misclassifications occur between neighboring classes (for example, Overweight Level I vs. Overweight Level II) or between more distant classes (e.g., underweight misclassified as obesity). In obesity classification, adjacent categories often share overlapping anthropometric and behavioral patterns; therefore, confusion between adjacent classes is expected and generally considered "mild" errors. Conversely, confusion between distant classes would indicate poor discriminative ability. By incorporating the confusion matrix heatmap, the study provides stronger evidence that the ensemble method performs reliably in a multi-class setting and that the remaining errors can be interpreted meaningfully.

As shown in Figure 5, most predictions are concentrated along the diagonal, while the remaining misclassifications appear in a few off-diagonal cells, indicating that errors mainly occur between closely related (adjacent) obesity categories rather than distant classes.

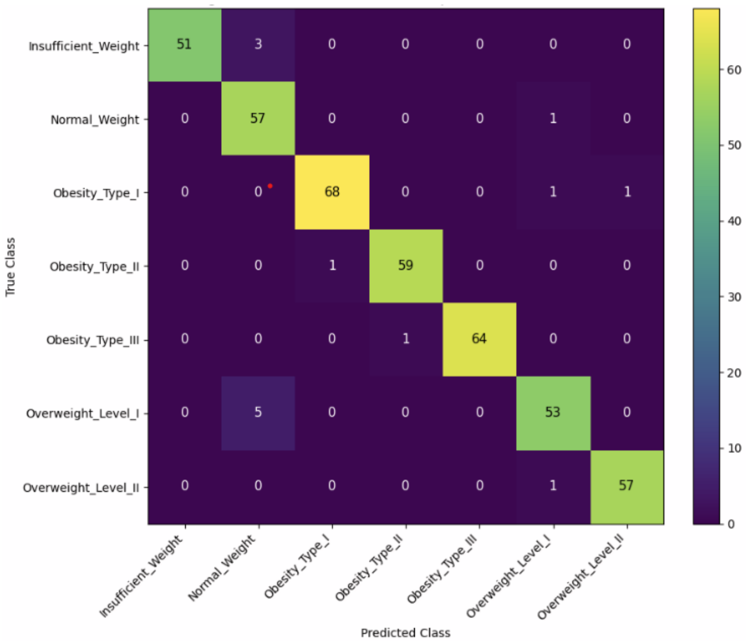

Figure 5. Confusion Matrix Heatmap of the Ensemble Model

3.7 Accuracy Comparison and Interpretation of Performance Improvements

Figure 6 compares the accuracies of the three approaches: LightGBM (96.45%), CatBoost (95.74%), and the ensemble model (96.69%). The visualization demonstrates that all models achieved high accuracy, but the ensemble model produced the best overall results. This confirms that combining boosting models with different tree structures and learning strategies can enhance classification performance. Importantly, the ensemble improvement is accompanied by increased macro and weighted average values, indicating better balance across classes and not merely improved performance in dominant categories.

In addition, the ensemble approach is more stable because it aggregates decision patterns from two distinct learning mechanisms. LightGBM may perform exceptionally well on certain patterns due to its leaf-wise optimization, while CatBoost may perform better on cases influenced heavily by categorical behaviors. Averaging probabilities tends to "smooth" predictions, reducing the risk that one model's overconfident wrong prediction determines the final class. This explains why the ensemble not only improves accuracy but also improves consistency and robustness.

As shown in Figure 6, the ensemble model achieves the highest accuracy (96.69%), slightly outperforming LightGBM (96.45%) and CatBoost (95.74%), which supports the benefit of combining both models.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

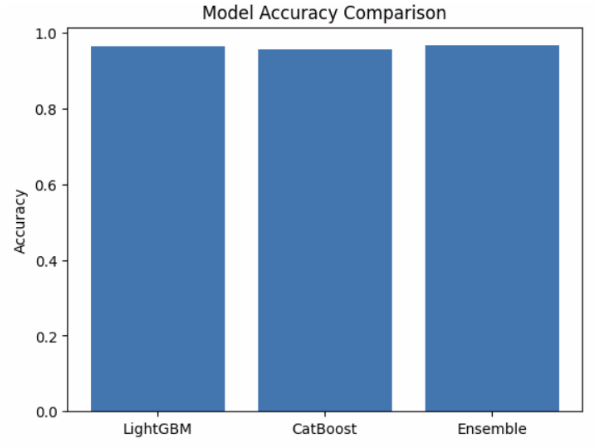

Figure 6. Accuracy Comparison

3.7 Practical Implications and Relation to the Research Problem

The results demonstrate that the proposed LightGBM–CatBoost ensemble method effectively solves the research problem by producing a high-performing and stable obesity risk classification model for productive-age individuals. The model's high accuracy and balanced class-wise metrics indicate that it can identify obesity risk categories reliably using complex, multidimensional inputs. Such a model has potential practical value in early risk screening, public health monitoring, and decision-support applications. In particular, an AI-based classification approach can assist in identifying individuals at risk before obesity-related complications develop, enabling preventive interventions.

At the same time, it is important to interpret the findings within the study scope. The dataset used is publicly available and synthetic/balanced, which is useful for benchmarking but may differ from real clinical data in terms of noise, missing values, measurement errors, and population diversity. Therefore, while the results are strong and demonstrate the effectiveness of the proposed method, future work should validate the model using real-world clinical datasets, larger populations, and external testing scenarios. Nonetheless, within the context of the current dataset and experimental design, the results strongly support the conclusion that ensemble learning improves stability and accuracy compared to single-model approaches in multi-class obesity risk classification.

4. CONCLUSION

This study addresses the main problem in obesity risk classification among individuals of productive age, namely, how to develop a predictive model capable of handling complex data characteristics that include lifestyle factors, eating habits, physical activity, and multiple categorical variables. Based on the experimental results, it can be concluded that the ensemble-based machine learning approach combining LightGBM and CatBoost provides an effective solution to this problem. The evaluation results demonstrate that the LightGBM–CatBoost ensemble model achieved an accuracy of 96.69%, with consistently high precision, recall, and F1-score values across all obesity risk classes. All classes obtained F1-scores above 0.93, with several classes achieving near-perfect performance, indicating that the model performs well not only globally but also in a balanced manner across different risk categories. The macro average and weighted average values of 0.97 further confirm the stability and reliability of the proposed model. Compared to the individual LightGBM and CatBoost models, the ensemble approach consistently produced superior performance. The probability averaging technique enables the integration of complementary strengths from both algorithms, where LightGBM effectively captures complex interactions among numerical features with high computational efficiency, while CatBoost excels in handling categorical variables and reducing overfitting. As a result, the ensemble approach minimizes the limitations of single models and improves overall prediction robustness. Despite these promising results, it should be noted that this study was conducted using a publicly available Kaggle dataset that is synthetic and balanced in nature. Consequently, the generalizability of the proposed model to real-world clinical or population-based data has not yet been fully validated. Future studies are therefore recommended to evaluate the model using real clinical datasets, larger and more diverse populations, and additional external validation scenarios. Nevertheless, the findings of this study demonstrate that the LightGBM–CatBoost ensemble method is a reliable and effective approach for obesity risk classification and has strong potential to serve as a foundation for data-driven decision support systems in the health sector, particularly for early obesity risk detection and prevention.

REFERENCES

[1] T. Aziz, N. Hussain, Z. Hameed, and L. Lin, "Elucidating the role of diet in maintaining gut health to reduce the risk of obesity, cardiovascular and other age-related inflammatory diseases : recent challenges and future recommendations," *Gut Microbes*, vol. 16, no. 1, 2024, doi: 10.1080/19490976.2023.2297864.
[2] E. Nunan *et al.*, "Obesity as a premature aging phenotype - implications for sarcopenic obesity," *GeroScience*, vol. 44, no. 3, pp.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 1, December 2025 | Hal 531-538
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i1.930

1393–1405, 2022, doi: 10.1007/s11357-022-00567-7.

[3] G. J. M. Yong *et al.*, "Precocious infant fecal microbiome promotes enterocyte barrier dysfuction , altered neuroendocrine signaling and associates with increased childhood obesity risk," *Gut Microbes*, vol. 16, no. 1, 2024, doi: 10.1080/19490976.2023.2290661.

[4] M. Ishida *et al.*, "The association between obesity, health service use, and work productivity in Australia : a cross - sectional quantile regression analysis," *Sci. Rep.*, vol. 13, no. 1, pp. 1–9, 2023, doi: 10.1038/s41598-023-33389-4.

[5] V. J. Beltrán-carrillo, Á. Megías, and D. González-cutre, "Elements behind sedentary lifestyles and unhealthy eating habits in individuals with severe obesity," *Int. J. Qual. Stud. Health Well-being*, vol. 17, no. 1, 2022, doi: 10.1080/17482631.2022.2056967.

[6] D. Mosha *et al.*, "Risk factors for overweight and obesity among women of reproductive age in Dar es Salaam, Tanzania," *BMC Nutr.*, vol. 7, no. 1, pp. 1–10, 2021, doi: 10.1186/s40795-021-00445-z.

[7] N. Opel *et al.*, "Brain structural abnormalities in obesity : relation to age, genetic risk, and common psychiatric disorders," *Mol. Psychiatry*, vol. 26, no. 9, pp. 4839–4852, 2021, doi: 10.1038/s41380-020-0774-9.

[8] A. Okunogbe, R. Nugent, G. Spencer, J. Powis, J. Ralston, and J. Wilding, "Economic impacts of overweight and obesity : current and future estimates for 161 countries," *BMJ Glob. Heal.*, vol. 7, no. 9, pp. 1–17, 2022, doi: 10.1136/bmjgh-2022-009773.

[9] R. Dettoni, C. Bahamondes, C. Yevenes, C. Cespedes, and J. Espinosa, "The effect of obesity on chronic diseases in USA: a flexible copula approach," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-28920-6.

[10] A. Bartosiewicz, J. Wyszyńska, P. Matłosz, E. Łuszczki, Ł. Oleksy, and A. Stolarczyk, "Prevalence of dyslipidaemia within Polish nurses. Cross-sectional study - single and multiple linear regression models and ROC analysis," *BMC Public Health*, vol. 24, no. 1, pp. 1–11, 2024, doi: 10.1186/s12889-024-18542-6.

[11] B. Zhang, D. Jiang, H. Ma, and H. Liu, "Association between triglyceride-glucose index and its obesity indicators with hypertension in postmenopausal women: a cross-sectional study," *Front. Nutr.*, vol. 12, pp. 1–10, 2025, doi: 10.3389/fnut.2025.1623697.

[12] S. Hamoud, A. Id, and L. Tafakori, "Predicting age at onset of childhood obesity using regression, Random Forest, Decision Tree, and K-Nearest Neighbour - A case study in Saudi Arabia," *PLoS One*, vol. 19, no. 9, pp. 1–21, 2024, doi: 10.1371/journal.pone.0308408.

[13] J. Kim, S. Mun, S. Lee, K. Jeong, and Y. Baek, "Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea," *BMC Public Health*, vol. 22, no. 1, pp. 1–10, 2022, doi: 10.1186/s12889-022-13131-x.

[14] A. Dwi, Y. Nur, and Y. Pristyanto, "Stock Price Time Series Data Forecasting Using the Light Gradient Boosting Machine (LightGBM) Model," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2270–2279, 2023, doi: 10.62527/joiv.7.4.1740.

[15] C. Zhang, J. Deng, and W. Yi, "Data-driven online tracking filter architecture: A LightGBM implementation," *Signal Processing*, vol. 221, p. 109477, 2024, doi: 10.1016/j.sigpro.2024.109477.

[16] S. Hussain, M. Wazir, T. A. Jumani, and S. Khan, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," *Energy Reports*, vol. 7, pp. 4425–4436, 2021, doi: 10.1016/j.egyr.2021.07.008.

[17] J. Dutta and S. Roy, "OccupancySense: Context-based indoor occupancy detection & prediction using CatBoost model," *Appl. Soft Comput.*, vol. 119, p. 108536, 2022, doi: 10.1016/j.asoc.2022.108536.

[18] I. D. Mienye, Y. Sun, and S. Member, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[19] T. Toharudin *et al.*, "Boosting Algorithm to Handle Unbalanced Classification of PM2.5 Concentration Levels by Observing Meteorological Parameters in Jakarta-Indonesia Using AdaBoost, XGBoost, CatBoost, and LightGBM," *IEEE Access*, vol. 11, pp. 35680–35696, 2023, doi: 10.1109/ACCESS.2023.3265019.

[20] R. P. Sari, F. Febriyanto, and A. C. Adi, "Analysis Implementation of the Ensemble Algorithm in Predicting Customer Churn in Telco Data: A Comparative Study," *Informatica*, vol. 47, no. 7, pp. 63–70, 2023, doi: 10.31449/inf.v47i7.4797.

[21] K. Shanmugavadivel, M. D. M. S, T. R. Mahesh, T. Al Shehari, and N. A. Alsadhan, "Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, pp. 1–22, 2024, doi: 10.1186/s12911-024-02688-9.

[22] M. Saber *et al.*, "Enhancing flood risk assessment through integration of ensemble learning approaches and physical-based hydrological modeling," *Geomatics, Nat. Hazards Risk*, vol. 14, no. 1, 2023, doi: 10.1080/19475705.2023.2203798.