Body Fat Analysis using Machine Learning Techniques

Ardra Nirwan[1,a)], Bhumi[1,b)], Gargi Bhardwaj[1,c)], Tarushi[1,d)] ,
Abhiruchi Sarswat[1,e)] and Poonam Bansal[1,f)]

[1]*Indira Gandhi Delhi Technical University for Women, New Delhi, India*

Author Emails
a) *ardra025bteceai24@igdtuw.ac.in*
b)*bhumikk0087@gmail.com*
c)*gargi126btcsai21@igdtuw.ac.in*
d)*tarushi143btcsai21@igdtuw.ac.in*
e)*poonambansal@igdtuw.ac.in*
f) *abhiruchi090btcsai21@igdtuw.ac.in*

Abstract: Body Fat Percentage (BFP) may serve as a more accurate indicator of health risks compared to traditional metrics such as Body Mass Index (BMI). However, conventional methods for measuring BFP like DEXA scans, hydrostatic weighing, and bioelectrical impedance are often costly, invasive, and not readily accessible to the general public. This research aims to develop an efficient and non-invasive method for predicting body fat percentage by employing machine learning techniques based on anthropometric measurements. Exploratory Data Analysis (EDA) was performed to understand feature distributions and relationships. The dataset underwent preprocessing, which included outlier removal, normalization, and division into training and testing sets. Five models were trained and evaluated: Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Gradient Boosting Regressor. The performance of each model was assessed using key metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score. Among the models, the Random Forest Regressor demonstrated the highest accuracy, achieving an R² of 0.9985. Comparative visualizations and an Actual vs. Predicted plot further validated the effectiveness of the approach. This study not only identifies significant predictors of body fat but also proposes an accessible solution for BFP estimation, with practical implications for healthcare, fitness, and weight management.

Keywords: Body Fat Prediction, Machine Learning in Healthcare, Regression Models, Model Evaluation

INTRODUCTION

In clinical settings, as well as in public health and the broader health and fitness sectors, obesity is typically identified based on the body mass index (BMI) standard, where a BMI of 30 kg/m2 or higher classifies both males and females as obese [1]. Obesity is a metabolic condition marked by an excess of body fat and is linked to a heightened risk of hypertension, diabetes, coronary heart disease, and cancer, which are significant health issues in both developed and developing nations [2]. Different techniques, including underwater weighing, DEXA scans, bioelectrical impedance analysis (BIA), magnetic resonance imaging (MRI), air displacement plethysmography, and near-infrared interactance, are employed to evaluate body fat. However, these methods generally require specialized and expensive equipment [3]. An alternative to Dual-Energy X-ray Absorptiometry (DXA) is Bioelectrical Impedance Analysis (BIA), which is relatively inexpensive and easy to implement. BIA can be utilized for hospitalized patients as well as in field studies, particularly in situations where there is a lack of specialized diagnostic equipment and trained medical staff available [4]. The rising prevalence of obesity in companion animals has led to a growing need for weight management programs. As a result, Body condition scoring (BCS) systems have gained widespread acceptance and are now commonly utilized to assess and monitor body fat levels [5]. New alternative anthropometric indices are being consistently developed to address the shortcomings of BMI. Among the most significant are the body shape index (ABSI), the body roundness index (BRI), and the body adiposity index (BAI) [6]. The associations of changes in predicted LBMI, ASMI, and BFMI with CVD risk were statistically significant in older or healthier women, possibly due to the higher number of CVD events among older women, as those in their 30s face greater CVD risk than those in their 20s [7]. A significant statistical challenge lies in examining the correlation between body fat and BMI. This includes analyzing the strength of the relationship, testing hypotheses regarding its functional structure, and applying the derived model for predictive purposes [8].

Recent non-invasive techniques for assessing fatty liver, such as MRI-PDFF, demonstrate high accuracy but come with a significant cost and abdominal ultrasound (US) is frequently suggested as the initial examination due to its accessibility and affordability, despite its limitations related to operator dependency and variability in results [9]. A cost-effective approach to more accurately predict adiposity involves developing an equation model for body fat percentage (BF%) derived from DXA, while primarily utilizing a less precise measure like BMI as the main predictor, along with additional covariates [10]. Research indicates that the Relative Insulin (RI) is a more accurate predictor of body fatness during puberty compared to Body Mass Index (BMI) [11]. Deep learning (DL) techniques are being utilized more frequently in quantitative susceptibility mapping (QSM). Prior research has demonstrated that convolutional DL networks can effectively address challenges such as dipole inversion and background field correction in QSM [12]. In recent years, a growing number of researchers have employed machine learning and data mining algorithms to diagnose and manage health issues such as heart and brain diseases. The non-invasive characteristics and precision of these technologies allow healthcare professionals to swiftly identify individuals at risk and to develop more effective preventive and management approaches [13]. Challenges such as incomplete data and limited sample sizes add complexity to the examination of these factors, which can be more effectively tackled through the application of machine learning methods, including support vector machines (SVMs) [14].

RELATED WORKS

The human body consists of muscle, lean tissues and fats, all of which must remain balanced. Excess fat can lead to obesity, which contributes to serious health issues. Therefore, reliable body fat prediction (BFP) is important for effective obesity treatment. Traditional BFP prediction models can be inaccurate and labor-intensive. Muhammed Kürşad Uçar [15] proposed a gender-based Body Fat Prediction value estimation model with ECG signals combined with machine learning algorithms.Tathiany J. Ferreira [16] demonstrated that AI-2D photo offers a practical, accessible, and user-friendly alternative to DXA for BFP estimation, supporting its use in modern clinical body composition analysis. Solaf A. Hussain, Nadire Cavus and Boran Sakeroglu [17] introduced a hybrid machine learning model combining Support Vector Regression (SVR) and Emotional Artificial Neural Networks (EANN) to predict body fat percentage (BFP) accurately. SVR was used to select key features, while EANN handled the prediction. Zongwen Fan and Jin Gou [18] proposed a cost-effective, fuzzy-weighted support vector machine model using simple body measurements. To handle noise, fuzzy weights and the Whale Optimization Algorithm (WOA) are used. Ivona Mitu and Manuela Ciocoiu [19] developed cost-effective ANN and MLR models to predict abdominal fat and fat-free mass using simple measurements. ANN outperformed MLR in accuracy and lower error rates. Waisbren and Eriksson [20] found that obesity measured by percent body fat (%BF) is a stronger predictor of surgical site infection (SSI) than body mass index (BMI).According to A. M. Spungen [21], body fat was evaluated using eight different techniques, with BIA, DEXA, TBW, and STK identified as dependable methods for estimating fat levels in individuals with tetraplegia evaluated the effectiveness of body mass index (BMI) compared to other body composition indexes and found that BMI-for-age is a more reliable screening tool than RI-for-age for identifying overweight and underweight conditions in children and adolescents. Jin-Tian Chen [22] demonstrated that the SVR model outperformed the ANN model in predicting both breast muscle and abdominal fat weights.

PROPOSED METHOD

1 Data Specification: For this study, we used the "Body Fat Prediction Dataset" dataset from Kaggle. The dataset comprises measurements from 252 adult males, featuring variables such as body density, percentage of body fat, age, weight, height, and various body circumference measurements (including neck, chest, abdomen, hip, and others)[23]. A structured overview of the dataset’s attributes along with their respective units is provided in Table 1, showcasing the features considered in the body fat prediction process.

S.No.	Attribute Name	S.No.	Attribute Name
1	Density (gm/cm³)	9	Hip (cm)
2	BodyFat (%)	10	Thigh (cm)

3	Age (Years)	11	Knee (cm)
4	Weight (lbs)	12	Ankle (cm)
5	Height (inches)	13	Biceps (cm)
6	Neck (cm)	14	Forearm (cm)
7	Chest	15	Wrist (cm)
8	Abdomen (cm)		

Table 1: Description of Attributes in the Body Fat Prediction Dataset

2 Data Preprocessing and Preparation: To begin the analysis, we explored the dataset to see how it was organized and whether there were any empty or missing parts. We found that 'Height' and 'Weight' were mixed up, the data correction was made by swapping the columns. Boxplots were used to detect outliers in each feature, and the IQR method was applied to remove them. An Extra Trees Regressor was used to determine the most important features for predicting body fat, highlighting the top five. A correlation heatmap was created to visualize how the variables are related, helping to better interpret the data and the model [24]. During the data preparation stage, the features were distinguished from the target variable, which was identified as 'BodyFat.' The dataset was subsequently divided into training and testing subsets, with 80% designated for training purposes and 20% reserved for testing. To maintain consistent model performance, particularly for algorithms that are sensitive to feature scaling, the StandardScaler was utilized to standardize the features, bringing them to a uniform scale. This process enhances the accuracy and convergence of machine learning models [24].

3 Models Training And Evaluation: Linear Regression ,Ridge Regression ,Lasso Regression, Random Forest Regressor were applied in this study. This study evaluated each model based on these key metrics:

- Mean Absolute Error (MAE): This metric gives the normal measure of the contrasts between anticipated and real values, overlooking whether the forecasts are over or beneath.
- Root Mean Squared Error (RMSE): RMSE takes the average of the squared prediction errors and then finds the square root of that average. It emphasizes larger errors more than smaller ones, making it useful for spotting models that perform poorly on certain predictions.
- R^2 Score (Accuracy): A value near 1 means the model closely aligns with the actual data, showing strong predictive ability.

Performance Comparison of Regression Models for Body Fat Prediction:
Table 2 outlines the evaluation results of different regression models, offering insights into their effectiveness through MAE, RMSE, and R^2 scores.

S. No.	Model Name	Mean Absolute Error	Root Mean Squared Error	R^2 Score (Accuracy)
1.	Linear Regression	0.4826	0.6215	0.9928
2.	Ridge Regression	0.5344	0.6737	0.9916
3.	Lasso Regression	0.5369	0.4469	0.9947
4.	Random Forest	0.2092	0.2836	0.9985

	Regressor			
5.	Gradient Boosting Regressor	0.2223	0.3446	0.9978

Table 2: Performance Comparison of Regression Models

4: Model Performance Visualization:

1. We used a scatter plot comparing actual body fat percentages to predicted values for evaluating the performance of the model visually.A red dashed line was included to indicate the threshold for perfect predictions. The closer the points are to this line, the better the model's accuracy. As illustrated in Figure 1, the model's accuracy is reflected by how closely the data points align with the reference line.

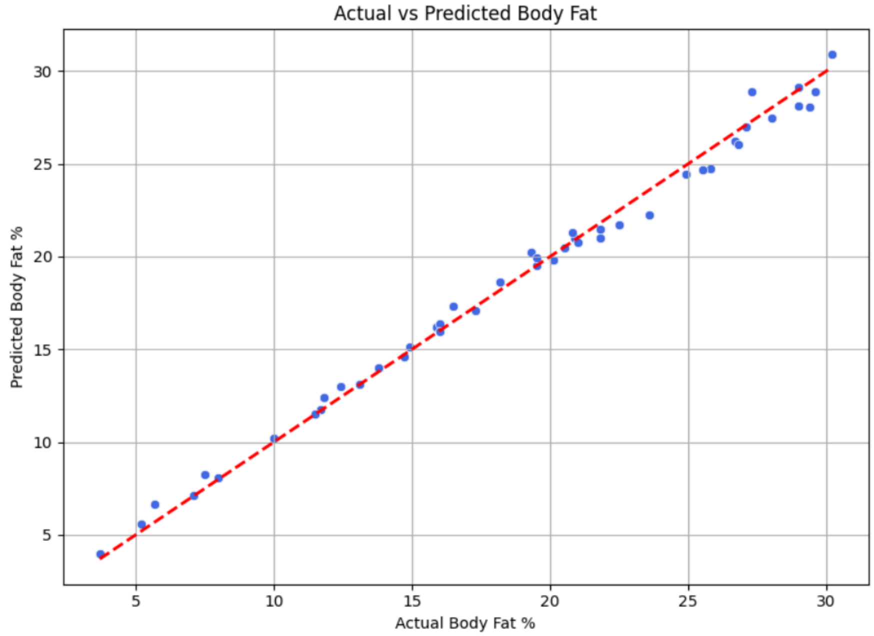

Fig 1: Scatter plot comparing actual BFP to predicted values

We generated a bar plot to evaluate the performance of various regression models, utilizing their R² scores for comparison. The models were arranged in descending order of accuracy, with the x-axis representing each model and the y-axis indicating their respective R² scores.

RESULTS AND DISCUSSIONS

A bar chart was used to compare the R² scores of five regression models, indicating their accuracy in predicting body fat percentage. Higher R² scores reflect better model performance. The chart clearly shows that the Random Forest Regressor attained the highest R² score of 0.9985, signifying that it accounted for nearly 99.85% of the variance in body fat percentage. This result suggests that the model's predictions were highly accurate, with minimal error in estimating body fat values. The Gradient Boosting Regressor followed closely, achieving an R² score of 0.9978, which affirms its strong predictive capabilities, albeit slightly less precise than the Random Forest model.

Lasso Regression demonstrated commendable performance with an R² score of 0.9947, indicating that it explained a substantial portion of the variance, though it was not as effective as the two tree-based models. Linear Regression and Ridge Regression, with R² scores of 0.9928 and 0.9916 (refer to Table:2) respectively, were the least effective models in this analysis. While they performed adequately, their predictive accuracy was inferior to that of the tree-based models, particularly the Random Forest and Gradient Boosting models. The findings show that machine learning models, particularly ensemble tree-based ones such as Random Forest and Gradient Boosting, are more effective than traditional regression methods when it comes to predicting body fat percentage. These models are great at identifying complex, non-linear relationships, which results in better accuracy. In contrast to Linear and Ridge Regression, tree-based models are better equipped to manage non-linearity, interactions between features, and noisy data. This underscores the promise of machine learning for providing more precise, non-invasive, and easily accessible body fat measurements, presenting a strong alternative to conventional techniques like DXA, BIA, and hydrostatic weighing.

CONCLUSION

Accurately estimating body fat percentage is crucial for assessing normal health, managing weight problems, and comparing fitness stages. While conventional strategies like hydrostatic weighing, dual-power X-ray absorptiometry (DXA), and bioelectrical impedance analysis (BIA) are precise, they tend to be expensive, time-consuming, and require specialized devices. In contrast, anthropometric-based predictive fashions present a non-invasive, cost-effective, and reachable alternative for estimating body fat percentage. In this study, we utilized system mastering strategies to develop a green predictive version that estimates frame fat using anthropometric features. We conducted sizable information preprocessing, which protected against lacking values, did away with outliers, and applied function scaling to ensure records of high quality. Exploratory data analysis (EDA) provided precious insights into feature distributions and correlations, which guided our selection of relevant predictors. To discover the key anthropometric variables that impact frame fat percent, we hired the greater bushes regressor for function choice. We educated and evaluated several device mastering fashions, inclusive of linear regression, random forest regressor, gradient boosting regressor, and help vector regression (SVR). We assessed overall performance using absolute error (MAE), implicit squared blunders (MSE), root imply squared blunders (RMSE), and the R² rating. The quality-performing model confirmed high accuracy in predicting frame fat percentage, outperforming traditional regression processes. This study underscores the effectiveness of system mastering in estimating frame rates and offers a strong framework for destiny research. Predicting body fat percentage using machine learning offers exciting opportunities in fitness and healthcare. Machine learning models provide a non-invasive, affordable alternative to methods like DEXA scans. Integrating data from fitness trackers and smartwatches improves accuracy through continuous monitoring and personalized feedback. Using diverse datasets makes the models more adaptable to various populations. Web and mobile apps can offer convenient, instant predictions. Deep learning techniques capture complex relationships for superior performance. Body fat estimates help identify risks of conditions such as cardiovascular disease and obesity. Real-time data and personalized guidance empower users to optimize fitness and nutrition for a healthy lifestyle. Machine learning-based body fat prediction offers a state-of-the-art, user-friendly, and effective solution for health monitoring and personalized fitness programs.

REFERENCES

1. Swainson, M. G., Batterham, A. M., Tsakirides, C., Rutherford, Z. H., & Hind, K. (2017).
2. Pongchaiyakul, C., Kosulwat, V., Rojroongwasinkul, N., Charoenkiatkul, S., Thepsuthammarat, K., Laopaiboon, M., ... & Rajatanavin, R. (2005).
3. Fan, Z., Chiong, R., & Chiong, F. (2022). A fuzzy-weighted Gaussian kernel-based machine learning approach for body fat prediction. Applied Intelligence, 52(3), 2359-2368.
4. Macek, P., Biskup, M., Terek-Derszniak, M., Stachura, M., Krol, H., Gozdz, S., & Zak, M. (2020). Optimal body fat percentage cut-off values in predicting the obesity-related cardiovascular risk factors: a cross-sectional cohort study. Diabetes, metabolic syndrome and obesity, 1587-1597.
5. Dugdale, A. H., Grove-White, D., Curtis, G. C., Harris, P. A., & Argo, C. M. (2012). Body condition scoring as a predictor of body fat in horses and ponies. The Veterinary Journal, 194(2), 173-178.

6. Yang, H. I., Cho, W., Ahn, K. Y., Shin, S. C., Kim, J. H., Yoo, S., ... & Jeon, J. Y. (2020). A new anthropometric index to predict percent body fat in young adults. Public health nutrition, 23(9), 1507-1514.
7. Kim, S. R., Lee, G., Choi, S., Oh, Y. H., Son, J. S., Park, M., & Park, S. M. (2022). Changes in predicted lean body mass, appendicular skeletal muscle mass, and body fat mass and cardiovascular disease. Journal of Cachexia, Sarcopenia and Muscle, 13(2), 1113-1123.
8. Mills, T. C., Gallagher, D., Wang, J., & Heshka, S. (2007). Modelling the relationship between body fat and the BMI. International journal of body composition research, 5(2), 73.
9. Chou, T. H., Yeh, H. J., Chang, C. C., Tang, J. H., Kao, W. Y., Su, I. C., ... & Su, E. C. Y. (2021). Deep learning for abdominal ultrasound: A computer-aided diagnostic system for the severity of fatty liver. Journal of the Chinese Medical Association, 84(9), 842-850.
10. Xu, S., Nianogo, R. A., Jaga, S., & Arah, O. A. (2023). Development and validation of a prediction equation for body fat percentage from measured BMI: a supervised machine learning approach. Scientific Reports, 13(1), 8010.
11. Hanspach, J., Bollmann, S., Grigo, J., Karius, A., Uder, M., & Laun, F. B. (2022). Deep learning–based quantitative susceptibility mapping (QSM) in the presence of fat using synthetically generated multi‐echo phase training data. Magnetic Resonance in Medicine, 88(4), 1548-1560.
12. Nematollahi, M. A., Jahangiri, S., Asadollahi, A., Salimi, M., Dehghan, A., Mashayekh, M., ... & Shariful Islam, S. M. (2023). Body composition predicts hypertension using machine learning methods: a cohort study. Scientific reports, 13(1), 6885.
13. Wong, A. Y., Harada, G., Lee, R., Gandhi, S. D., Dziedzic, A., Espinoza‐Orias, A., ... & Samartzis, D. (2021). Preoperative paraspinal neck muscle characteristics predict early onset adjacent segment degeneration in anterior cervical fusion patients: a machine‐learning modeling analysis. Journal of Orthopaedic Research®, 39(8), 1732-1744.
14. Uçar, M. K., Ucar, Z., Uçar, K., Akman, M., & Bozkurt, M. R. (2021). Determination of body fat percentage by electrocardiography signal with gender based artificial intelligence. Biomedical Signal Processing and Control, 68, 102650.
15. Ferreira, T. J., Salvador, I. C., Pessanha, C. R., da Silva, R. R., Pereira, A. D., Horst, M. A., ... & Pierucci, A. P. (2025). Advances in the estimation of body fat percentage using an artificial intelligence 2D-photo method. npj Digital Medicine, 8(1), 43.
16. Hussain, S. A., Cavus, N., & Sekeroglu, B. (2021). Hybrid machine learning model for body fat percentage prediction based on support vector regression and emotional artificial neural networks. Applied Sciences, 11(21), 9797.
17. Fan, Z., & Gou, J. (2023). Predicting body fat using a novel fuzzy-weighted approach optimized by the whale optimization algorithm. Expert Systems with Applications, 217, 119558.
18. Mitu, I., Dimitriu, C.-D., Mitu, O., Preda, C., Mitu, F., & Ciocoiu, M. (2023). Artificial Neural Network Models for Accurate Predictions of Fat-Free and Fat Masses, Using Easy-to-Measure Anthropometric Parameters. Biomedicines, 11(2), 489. https://doi.org/10.3390/biomedicines11020489
19. Waisbren, E., Rosen, H., Bader, A. M., Lipsitz, S. R., Rogers Jr, S. O., & Eriksson, E. (2010). Percent body fat and prediction of surgical site infection. Journal of the American College of Surgeons, 210(4), 381-389.
20. Spungen, A. M., Bauman, W. A., Wang, J., & Pierson, R. N. (1995). Measurement of body fat in individuals with tetraplegia: a comparison of eight clinical methods. Spinal Cord, 33(7), 402-408.
21. Mei, Z., Grummer-Strawn, L. M., Pietrobelli, A., Goulding, A., Goran, M. I., & Dietz, W. H. (2002). Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *The American journal of clinical nutrition*, *75*(6), 978-985.
22. Chen, J. T., He, P. G., Jiang, J. S., Yang, Y. F., Wang, S. Y., Pan, C. H., ... & Pan, J. M. (2023). In vivo prediction of abdominal fat and breast muscle in broiler chicken using live body measurements based on machine learning. Poultry Science, 102
23. Gupta, S., Bhardwaj, G., Arora, M., Rani, R., Bansal, P., & Kumar, R. (2023, March). Employee Attrition Prediction in Industries using Machine Learning Algorithms. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 945-950). IEEE.
24. Jaiswal, G., Bhardwaj, G., Tarushi, Sarswat, A., & Rani, R. (2024). Predictive Modeling to Identify Syndrome Patterns. In *Healthcare Industry Assessment: Analyzing Risks, Security, and Reliability* (pp. 67-91). Cham: Springer Nature Switzerland.
25. Sarswat, A., Bansal, P., & Malik, K. (2023, January). Analysis of heart failure survival rates in middle-aged and older people using machine learning model. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. i-viii). IEEE.