Prediction of Body Fat Percentage Based on Anthropometric Measurements Using Data Mining Approach

Hamsa Amro[1], Mohammed Awad[*2]

[1]Health Science Department, AAUP, Oral and Dental Health Unit, MOH, Ramallah- Palestine
h.amro@student.aapu.edu
[2]CSE Department, Artificial Intelligence, Faculty of Eng. & IT, Arab American University-Palestine
Mohammed.awad@aaup.edu

Abstract

In recent years, heart disease, diabetes, and some types of cancers have been reported as some main causes of death in most countries of the world, and obesity, which is often attributed to excess body fat, is one of the most common risk factors for these diseases. To make the vast amounts of data produced by health care information systems useful to the potential, the researchers applied knowledge discovery through predictive modeling. This study used anthropometric measurements as input data to different data mining techniques to predict body fat percentage. Fisher's Method of Scoring was used to select the most effective features in the prediction process, which were represented by eight variables (abdomen, BMI, chest hip, neck, thigh, knee, and age). The selected features were used as input\output for data mining approaches like multiple regressions (MR), Support Vector Machine (SVM), and Artificial Neural Networks (ANNs). As a result, ANNs outperformed the other methods used to predict body fat percentage by correlation coefficient R^2=0.77 with eight selected anthropometric parameters. This outperformance was attributed to ANNs high ability to predict and their understandable and implementable resulting knowledge. Therefore, the researchers concluded that ANNs could be used and applied to a dataset of the Palestinian population.

keywords: Obesity, Correlation, Data mining, Artificial Neural Network, Prediction

Introduction

Obesity and related health risks have been observed to be epidemiological issues in most countries of the world (Abdeen et al., 2012). Within the Eastern Mediterranean Region, an increasing prevalence of overweight has been recorded as stated by the World Health Organization (WHO, 2020).It has been proven that the increase in the level of obesity among the population often precedes the increase in the occurrence of chronic diseases, such as high blood pressure and diabetes (Luke et al., 1997).

To measure and monitor obesity levels, easy and convenient methods of determining body composition and measurements have to be used. The most common way to determine the level of obesity is Body Mass Index (BMI) and Fat Body Percentage (FBP) (Kupusinac, Stokić, Sukić, Rankov, & Katić, 2017). BMI is a general indicator of nutritional status, which is defined as the percentage of body mass and square of the body height, while FBP is a better predictor of visceral fat mass, as the fat mass can be predicted using the two-compartment equation developed by Siri based on the assumed densities of both fat mass and lean mass . Therefore, BFP can be determined from the calculated body density (“SOCR Data BMI Regression - Socr,” n.d.).

Despite the common belief amongst people that BMI is the gold standard for measuring the level of obesity, BMI is not a measure of the fat mass in the body; the aging period is characterized by reducing body height and increasing fat mass through the redistribution of fat in the form of visceral deposits, even if BMI and body weight is maintained (Kupusinac et al., 2017). Therefore, it will be important if BFP can be predicted from easily measurable parameters, such as anthropometric and age data away from the traditional methods that are characterized by calculating the individual body shape without taking any other considerations, and thus does not provide an accurate insight into a prediction of body composition. Hence, other more accurate BFP prediction methods have to be developed without the need for expensive equipment (Uçar, Uçar, Köksal, & Daldal, 2021).

These modern prediction methods are represented by using data mining techniques to help diagnose and predict diseases that may threaten human life. Data mining and machine learning techniques are gaining great importance in the health sector as in other fields. They facilitate the analysis and use health-related data sets to provide more efficient health care (AlAgha, Faris, Hammo, & Al-Zoubi, 2018).

The current research aimed to analyze the regression capacity for Machine Learning (ML) and investigate its accuracy through Multiple Regressions (MR) and Vector Support Machine (SVM) (alabdallah, & awad, 2018) using anthropometric body measurements to predict BFP. The researcher also included a comparison with the results obtained by the Artificial Neural Networks (ANNs) (Hamdan. et al., 2018) since these models have recently been widely applied to machine learning problems

Literature Review

Overweight and obesity management remains a major challenge in managing public health worldwide(Babajide et al., 2020). In developing countries, weight gain is concentrated in low-income groups(Ford, Patel, & Narayan, 2017). There are many reports in Palestine indicating an increase in the level of Non-Communicable Diseases (NCDs), such as high blood pressure, diabetes, and heart disease, which are directly related to obesity and overweight; death related to NCDs has also increased. In Palestine, there are very few studies that deal with overweight and obesity. The prevalence of obesity has been reported by the World Health Organization to be as high as 26.8 % of the Palestinian population (WHO, 2015).

Numerous research and projects have been carried out on the prognosis of obesity-related risks by predicting the BFP using data mining. Jindal et al**.** (2018) performed ensemble machine learning approaches to predict obesity based on key determinants of height, weight, age, as well as BMI. The ensemble model utilized a generalized linear model, Random Forest (RF), and partial least square, with 90% prediction accuracy. Kupusinac et al. (2014) applied a new approach from the ANNs by using determinants of gender, age, and BMI to a sample that included 2755 randomly selected cases from the northern part of Serbia. In terms of accuracy, the developed ANNs approach outperformed the previously used rule-based methods on the same determinants (Deurenberg et al.). Shoa (2014) proposed hybrid models capable of predicting the effectiveness of BFP, such as the Multiple Regression Model (MRM), Artificial Neural Network (ANNs), Multivariate Adaptive Regression Splines (MARS), and Support Vector Regression (SVR) techniques. These models were applied to a real dataset by Johnson (1996) (Johnson & College, 1996). The proposed hybrid models with fewer relevant variables outperformed the single prediction models except for the ANN hybrid models whose percentage improvements of the Mean Absolute Percentage Error (MAPE) for the proposed MR-MARS model over the single MR and MARS models indicated 4.2% and 9.5%, respectively.

While Merrill et al. (2020a) developed a statistical regression model through a study conducted on 228 adults aged between 21 and 70 years, where the model was applied to a training set that contained 163 records, which included data on age, anthropometry measures, body mass index, and skinfold measurements. As a result, the developed regression model outperformed the other four pre-existing methods (Durnin, Hodgdon, Jackson, and Woolcott) and had an average error of less than 0.10%. Uçar et al (2021) conducted a study on a real data set consisting of 13 anthropometric measurements for male adults, similar to Shao's and the BFP was determined using hybrid machine learning methods, Multilayer Feed Forward Neural Networks (MLFFNNs), and Support Vector Machine Regression Model (SVMs), and Decision Tree regression model (DT) was also developed with a high accuracy rate and a minimum of parameters. The results showed that the hybrid system can be used to estimate BFP with only one anthropometry with a correlation value of R= 0.79.

Through our study, it is possible to investigate the ability of ML regression by developing a regression model of Support Vector Machines (SVMs) and Multiple Regression (MR) using fewer relevant BFP determinants such as age, body mass, and abdominal circumference, and comparing the results with those obtained by Artificial Neural Networks (ANNs) without using hybrid models

Materials and Methods

In order to investigate the ability of regression for ML and ANNs in predicting BFP, a real data set, which was originally from Human Performance Research Center, Brigham Young University, and provided by Dr. A. Garth Fisher in 1994, was analyzed. Body Fat Percentage, age, weight, height, and ten body circumference measurements were recorded for 252 men. In this study, we used Matlab software to clean and convert data set inputs for the purpose of mining and evaluating the output. In the feature selection phase, all features were selected according to their correlation values. In the prediction phase, a machine learning model, such as (Multiple Regression Model, Support Vector Machine) and Levenberg-Marquardt algorithm (Awad, & Zaid-alkelani, 2019) were used to propose an artificial neural network approach to predict obesity.

BFP Dataset

A real data set was analyzed to estimate BFP. The dataset included 13 anthropometric measurements and BFP values for 252 adult males. Table 1 shows a description of the measurements used, and table 2 shows the important circumference dictionary.

Due to the difficulty in measuring BFP, the Siri equation was used to calculate the BFP value (Johnson & College, 1996)

$$Y = \frac{495}{D} - 450 \qquad (1)$$

Where **Y** denotes the BFP and **D** denotes Density determined from underwater weighing.

Table 1: Variables' definition in the BFP dataset.

Variable	Meaning
D	Density determined from underwater weighing
Y	BFP
X1	Age (years)
X2	Height (cm)
X3	Weight (kg)
X4	Neck circumference (cm)
X5	Chest circumference (cm)
X6	Abdomen 2 circumference (cm)
X7	Hip circumference (cm)
X8	Thigh circumference (cm)
X9	Knee circumference (cm)
X10	Ankle circumference (cm)
X11	Biceps (extended) circumference (cm)
X12	Forearm circumference (cm)
X13	Wrist circumference (cm)

Table 2: Circumferences dictionary

Anthropometrical measurement	Definition of circumferences
Neck	The level just below the laryngeal prominence perpendicular to the long axis of the neck.
Chest	The level of the nipples or at the level just below the scapular angles
Abdomen	The midpoint of the line between the rib or costal margin and the iliac crest in the midaxillary line
Hip	The distance around the human body at the level of the maximum posterior extension of the buttocks
Thigh	The midpoint between the inguinal crease and the proximal border of the patella
Biceps	Taken at the largest part of the bicep.

Data Pre-processing

In the mentioned dataset, after calculating the BFP by the Siri equation, the density variables were deleted. Measurements of height and weight in the data set were measured in inches and pounds; the unit of measurement of length was converted from inch into cm, where 1 inch = 2.54 cm, and the unit of weight measurement was converted from pound into kg, where 1 pound = 0.453 kg, while the anthropometric measurements were all in centimeters. The study omitted each of the cases 96, 172, and 182 due to the deviation of the measurements of BFP from normal conditions, as it recorded 0, 0.7, and 0.3 respectively. While the height in case 42, and the weight in case 39 were indicated as outliers' values. Thus, they were omitted. See Fig. (1), Fig (2), and Fig (3). The sample size, after being treated, became 247 cases with 13 anthropometric measurements.

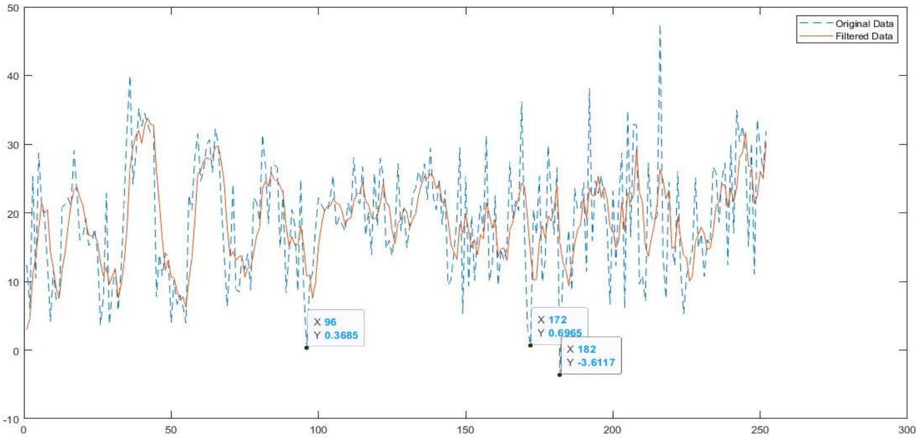


Figure 1: Body Fat Percentage filtration.

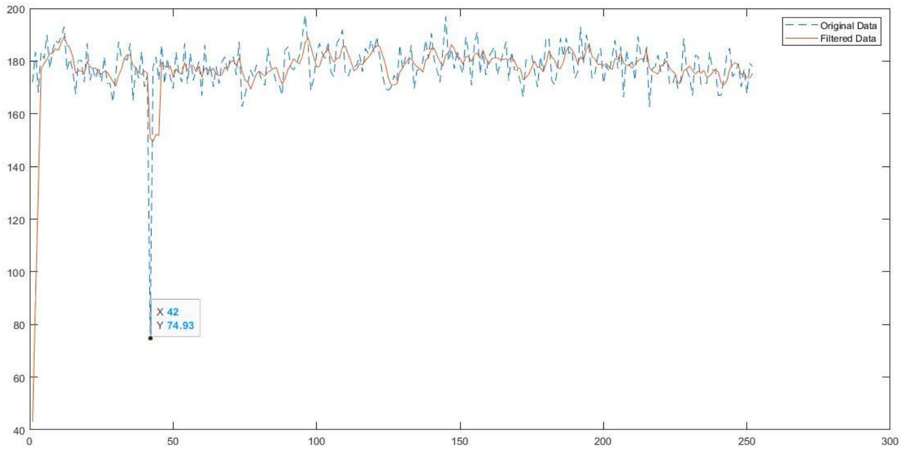


Figure 2: Height filtration.

Feature Selection and Data Enriching

Body mass index (BMI) was derived from the height and weight variables according to the following equation:

$$Weight\ Kg/[height\ (cm)]^2 * 10{,}000 \quad (2)$$

The feature selection approach can simplify models and reduce data; which makes the results understandable. Defining a subset of data by identifying the most important relevant features and removing a lot of redundant data will increase the efficiency and speed of the learning algorithms. The 12 explanatory variables were reduced to eight according to their importance through the use of Fisher Score (table 3); which is one of the most commonly supervised methods used for selecting features. However, each feature (Xi) is independently determined according to its great relationship with the target variable (BFP). The score of the i_{th} feature S_i was calculated by Fisher Score as shown in the following equation:

$$S_i = \frac{\sum n_j\,(\mu ij - \mu\_i)^2}{\sum n_{j}{*}p_{ij}^2} \qquad (3)$$

Where μ_{ij} and ρ_{ij} are the mean and the variance of the i_{th} feature in the j_{th} class, n_j is the number of inputs in the j_{th} class and μ_i is the mean of the $\underline{\mathrm{i}_{\mathrm{th}}}$ feature.

Table 3: Feature selection according to Fisher score.

Variable	Fisher score value
Abdomen	9.789757
BMI	7.703632
Chest	5.524361
Hip	5.035753
Neck	4.007346
Thigh	3.760747
Knee	3.368902
Age	3.16745
Ankle	3.018483
Biceps	2.93451
Wrist	2.542725
Forearm	2.020003

The eight features are: abdomen, BMI, chest, hip, neck, thigh, knee, and age.

Data Mining Algorithms

Based on previous studies, there are many kinds of prediction algorithm. The most three prominent data mining algorithms, which had the potential to achieve good results, are: Multiple Regression Model (MRM), Artificial Neural Networks (ANNs), and linear Support Vector Model (SVM) techniques.

Multiple Regression Model

The Multiple Regression Model predicts a numerical value. Regression performs operations on the given dataset, where the target values BFP are defined. The relationships that the regression establishes between the expectation values (body circumference measurements, BMI, and age) and target BFP can form a pattern, which can be used in other datasets to predict target values BFP. Multiple Linear Regression (MLR) is a model that allows us to study multiple effects variables. MLR model is identical to the simple linear regression model, with the only difference that there are more explanatory variables as shown in the following equation:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots + b_k \cdot x_k \qquad (4)$$

Where b is the mean of y when all xi is zero (when xi = 0 does not make sense, xi change in the mean of Y when Xi increases one unit while the others remain constant. (Wang. et al. 2019)

Support Vector Machine

Support Vector Machine (SVM) is one of the most powerful supervised theoretical machine learning algorithms as it is based on the Vapnik-Chervonenkis theory in calculating the linear regression function, where the regression modeling obtains the coefficients through minimizing the square error. SVM has been applied to various fields including medical diagnosis. It depends on finding the maximum geometrical margins between hyperplanes, which are the robustness of the SVM method, to determine the optimal hyperplane that separates the dataset classes in the middle of the maximum margin. To build a classification model of an n-dimensional dataset consisting of features, it needs an n-dimensional feature space. The sample vectors are separated by the optimal hyperplane into classes. A maximum margin between the support vectors, which are the closest points of both positive and negative classes to the optimal hyperplane, has to be taken into consideration. The maximum margins mean the minimum risk of misclassifying new data (Wang and Chen 2020).

In the SVM, the point is to maximize the margin between hyper-plane and point of data; the function that helps maximize the margin is called loss function c(x,y,f(x)) given by training data (xi, yi) for i = 1 ...N, with xi ∈ Rd and yi ∈ {−1, 1}, learn a classifier f(x).

$$c(x, y, f(x)) = \begin{cases} 0 \;\;, & if\ y * f(x) \geq 1 \\ 1 - y * f(x), & if\ y * f(x) < 1 \end{cases} \qquad (5)$$

$$c(x, y, f(x)) = (1 - y * f(x))_+ \qquad (6)$$

The cost equals zero. If the actual value and the predicted value are of the same sign, if the two values are not equal to zero, then it calculates the loss value by the cost function. When it adds a parameter called regularization parameters to the cost function, it looks as below, the regularization parameter aims to balance the margin maximization and loss by using (3.22), where Xi is the input, w is the weight and λ is regularization parameters:

$$\min_w \lambda \|w\|^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+ \qquad (7)$$

By using the loss function, it takes partial derivatives concerning the weights of the data points to find the gradients δthat able to update the weights represented by (3.23).

$$\frac{\delta}{\delta w_k} \lambda \|w\|^2 = 2\lambda w_k \qquad (8)$$

$$\frac{\delta}{\delta w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0 \;\;, & if\ y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_i, & if\ y_i \langle x_i, w \rangle < 1 \end{cases} \qquad (9)$$

No misclassification is in our method when correctly predicts the class of data points, it has one solution that is represented in (3.24) by updating the gradients from the regularization parameter.

$$\omega = \omega - \alpha.(2\lambda\omega) \qquad (10)$$

When our method makes an error when predicting the class of data points, this is called misclassification, it has one solution that is represented in (3.25) by updating the gradients including the loss along with the regularization parameter.

$$\omega = \omega + \alpha.(y_i.x_i - 2\lambda\omega) \qquad (11)$$

Artificial Neural Networks (ANNs)

Due to ANNs' characteristics of high noise tolerance and the ability to classify non-visual patterns, ANNs have been increasingly used for modeling non-static processes.

The nodes in ANNs can be divided into three layers: Input, Output, and one or more hidden layers (Walczak & Cerpa, 2003). The concept, on which artificial neural networks are based, is the simulation of the model to access data of these units for classification, prediction, analyzing, or any other treatment of input data. Neural networks have shown the capability of solving problems of predictions (Shahid, Rappon, & Berta, 2019). The neurons contained by the hidden layer determine the action to be made to the input data received from the input layer. In addition, the weights of these neurons, which are developed by the learning process, play a major role in the decision. Later, the data might be transferred to the output layer by applying the activation functions included in the neurons. Every input data has a related estimation of weight W. This esteem is a factor of significance since it updates itself within the training of the neural network with the goal that it shows signs of conduct improvement (Ahmed, 2005). The general structure of ANNs is illustrated as shown in Figure 3 below:

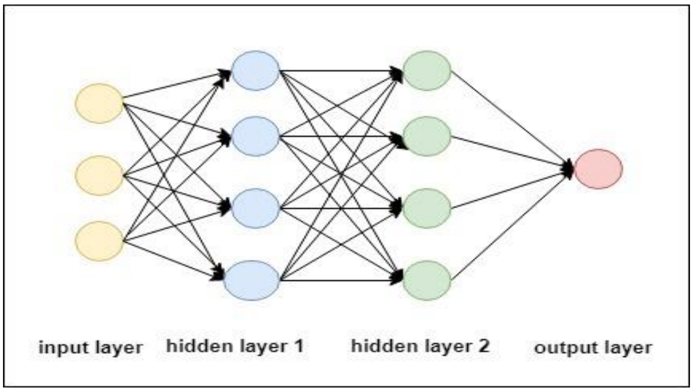


Figure 3: General model of multilayer artificial neural networks.

ANNs with feed-forward back-propagation algorithm consists of three layers represented by feed forward, multi-input multi-output as follows (Du, Y.2018):

- Input layer $X_{i,}$ i = 1, 2, …, n. Where n is the number of input nodes.
- Hidden layer j: Each node is a neuron, each neuron connected to the input layer by the processing unit called weights w_{ij}, where i is the input node and j is the hidden layer node.
- Output layer k: Contains the nodes that produce the output of the network represented by several neurons and depends on several outputs, Y_k.

When the training phase was fed to the input layer, the sum of weights from the input to the jth node in the hidden layer is given by:

$$y = \sum W_{ij} X_i + \theta_j \qquad (12)$$

θ_j: called the bias node that always has a value of 1; it calculates the gradients efficiently by back propagation algorithm when using ANNs. The back-propagation algorithm always starts from the last layer (output layer) and propagates backward to update the weights of the network; it needs an activation function, typically used the sigmoid function. The actual output of the jth node is:

$$Y_j = X_k = \frac{1}{1+e^{-y}} \qquad (13)$$

In the output layer, the difference value between the actual and the target value is Δ_k, where the actual value of the node k is Y_k and the target value is t_k, while X_k is the input to the next layer' node.

$$\Delta_k = t_k - Y_k \qquad (14)$$

δ_k: The error signal of the output layer is calculated by Δ_k and the derivative of the sigmoid function.

$$\delta_k = \Delta_k Y_k (1 - Y_k) \qquad (15)$$

The change in the weight between node j and node k is done by multiplying the error at node k by the output of node j by using the delta rule.

$$\Delta w_{jk} = l\, \delta_k X_k \qquad (16)$$

w_{jk}: The weight between node j and k, where l is the learning rate, so to update it by the following formula:

$$w_{jk} = w_{jk} + \Delta w_{jk} \qquad (17)$$

To calculate, the error signal $\delta_j$for node j in the hidden layer, δ_j: The error signal for node j in the hidden layer is calculated by the following formula:

$$\delta_j = (t_k - Y_k)Y_k \sum w_{jk}\, \delta_k \qquad (18)$$

w_{ij} is the weights between the input node i and the node j can be updated by using 14 and 15 so

$$\Delta w_{ij} = l\, \delta_j X_j \qquad (19)$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad (20)$$

The back-propagation algorithm repeats until the error on the output node is minimized.

Results

The predefined Matlab application was used to develop multiple linear regression for BFP prediction by fitting the observed data. Root mean square error (RMSE) on the validation set scores 4.4826 (lower RMSE is better than a higher one) and R-squared scores 0.70 (with 0 denoting that model does not explain any variation and 1 denoting that it perfectly explains the observed variation) for eight selected features. The score estimates the good performance of the trained model on the given data. For more details, see Figure 4 below.

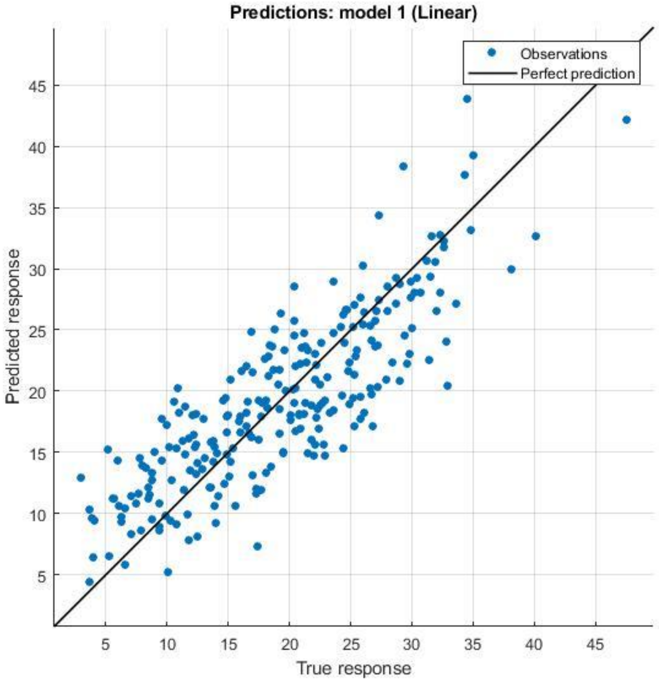


Figure4: MLR model for BFP prediction using eight independent variables.

This study also performed the linear SVM modeling of the BFP dataset. SVM aims to fit the curve in hyperplane so that it passes over the BFP dataset. The score of SVMs' RMSE indicates 4.5497 with a correlation coefficient R^2= 0.69. Figure 5 illustrates the residuals BFP in abdomen circumference.

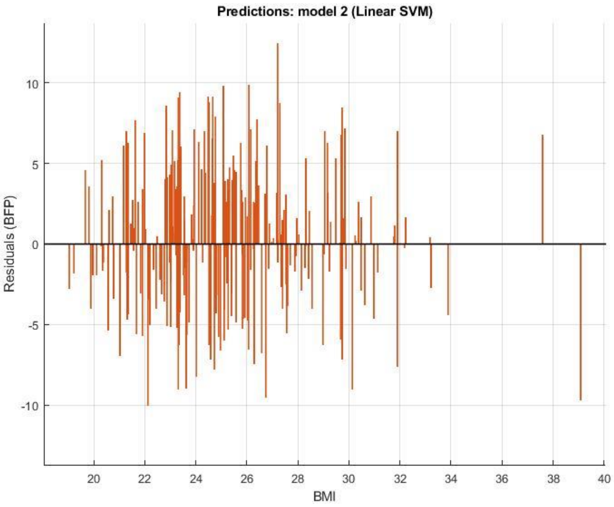


Figure 5: Linear SVM model for BFP prediction using BMI circumference.

According to an artificial neural network, the BFP dataset was divided into 70% for training (173 cases), and 30% for testing and validation (37, 37) cases.

The validation set is used to adjust model parameters, and the test set is used to evaluate performance. The (nnstart) function was used on the Matlab and ANN applied to ten hidden neurons, using Levenberg–Marquardt algorithm for solving generic curve-fitting problems.

The neural network was applied twice. First, each of the eight independent parameters was entered alone with BFP as an output. The results indicated that the Abdomen R=0.82 BMI (0.665), Hip (0.731 and Chest (0.674) (parameters had the highest correlation coefficient of more than R> 0.6 (the closer to 1, the better the training performance). While the other four parameters (Age, Neck, thigh, and knee) had a correlation coefficient of less than R<0.5, 0.14, 0.43, 0.56, 0.34 respectively. In the second time, the eight independent parameters were used as input to the neural network and BFP as an output, where training performance was excellent with correlation coefficient for both training, validation, and testing R= (0.89, 0.79, 0.77) respectively. See Figures 6, 7, and 8 for more details.

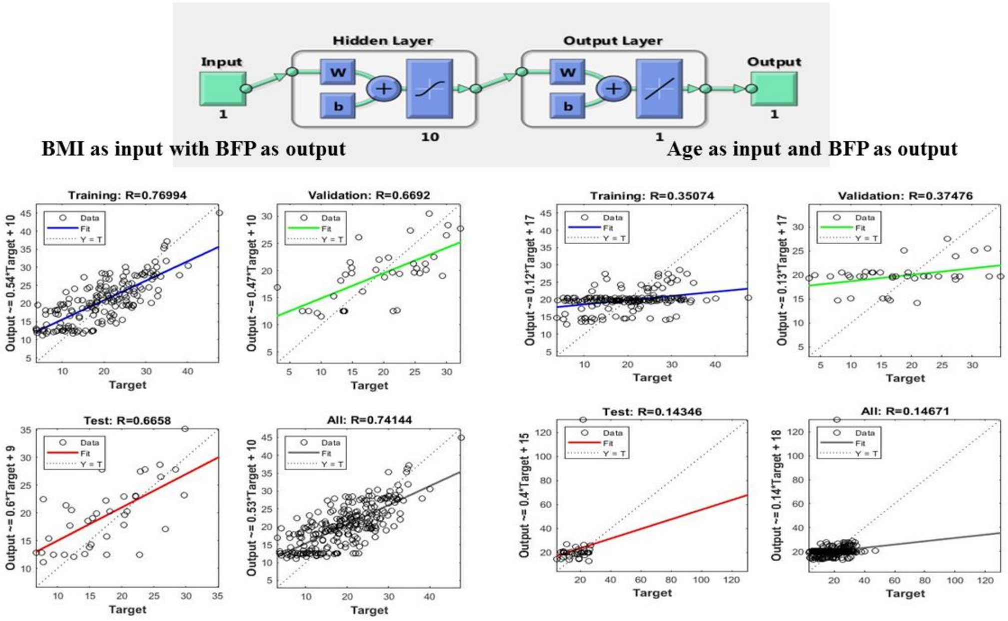
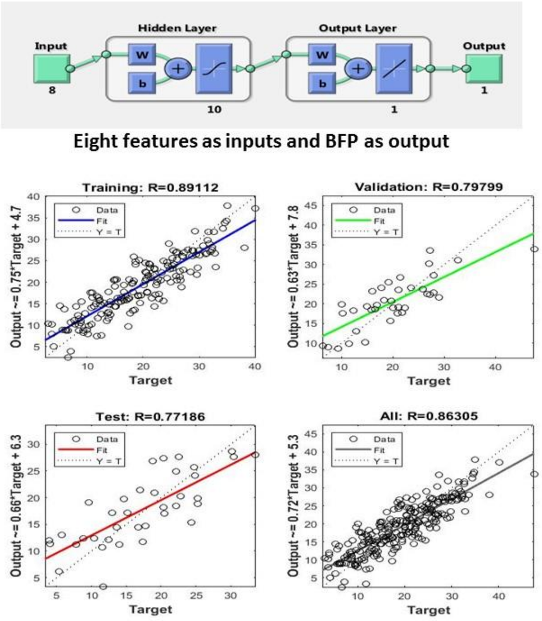


Figure 6: ANNs model for BFP prediction using one independent parameter.



Eight features as inputs and BFP as output

Figure 7: ANN model for BFP prediction using eight independents parameters.

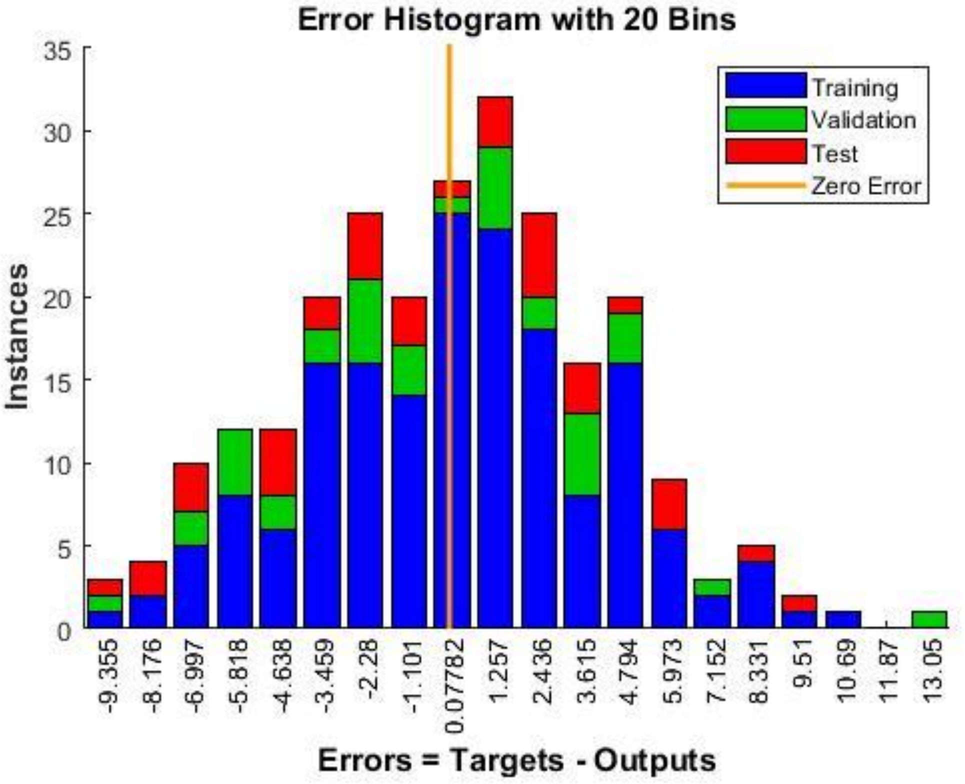
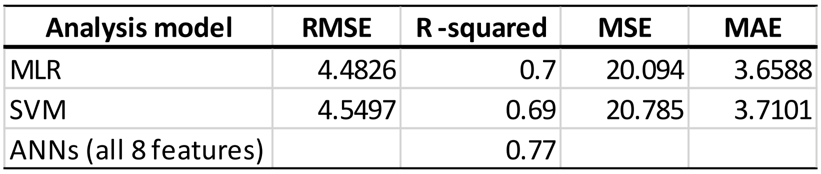


Figure 8: Error histogram for the ANNs model for BFP prediction using eight independents features.

The performance of the classifier is summarized in the table below

Table 4: Performance result analysis of classifier



Analysis model	RMSE	R -squared	MSE	MAE
MLR	4.4826	0.7	20.094	3.6588
SVM	4.5497	0.69	20.785	3.7101
ANNs (all 8 features)		0.77		

Discussion

The study aimed to predict the BFP using minimal data and a high accuracy rate using data mining algorithms, such as Multiple Linear Regression, Support Vector Machine, and Artificial Neural Networks. Compared to the performance of the algorithms used, (ANN) outperformed in constructing a prediction BFP model, followed by (SVM), then (MLR).

While developing models, a multiple regression model was created using the entire dataset. When whole data is used for training, the accuracy rate is high, which provides misleading information. The study models only focused on males, which inevitably led to worse R^2, but a more realistic estimate of how the model would work with a specific gender.

The ANNs failed to predict the BFP in four out of eight determinants when entered into the network separately; this confirms that the hybrid models can address these defects as indicated by Uçar et al. (Uçar et al., 2021)

Many published studies applied machine learning tools to predict BFP using anthropometric measurements, whereas our study has been distinguished by using fewer relevant BFP determinants and comparing the results, which were applied using machine learning tools, with the results that were applied using artificial neural networks without resorting to using hybrid models. Naturally, the study had several limitations represented by lacking relevant anthropometric measurements, such as waist circumference, and waist to hip ratio, which are important indicators of abdominal fat (Pereira et al., 2010)(Merrill, Chambers, & Cham, 2020b). The neural network algorithms had high accuracy despite their complexity; however, the prediction approach using them was influenced by the sample size, and consequently obtaining better performance in predicting BFP, and intelligence algorithms can be combined with genetic algorithms (Gao et al., 2020).

In this study, gender was not in consideration. Therefore, future comparative studies are recommended to predict BFP in both genders. In Palestine, obesity is one of the public health challenges facing the Palestinian population.

It is linked to non-communicable chronic diseases, such as diabetes, heart disease, and some types of cancer. The importance of health informatics has increased dramatically in recent years due to the need for safe and effective management of medical data (AlAgha et al., 2018). The health care field faces strong pressure to reduce costs while increasing the quality of the services provided, so it was necessary to highlight those systems that can extract patterns and knowledge from big data sets to support decision-making and knowledge management at low costs. (Almadhoun & El-Halees, 2017).

Conclusion

Since traditional methods of body composition prediction show low accuracy and poor adaptability, the researchers proposed a method for predicting body fat percentage through machine learning algorithms and artificial neural networks. In this method, preferred anthropometric parameters with the largest correlations with BFP were firstly obtained using Pearson Correlation Coefficient (r). Secondly, algorithms were developed using the MLR and SVM machine learning model, in addition to the development of the ANNs algorithm.

According to the modeling results, the developed ANNs approach outperformed other methods used to predict body fat percentage by correlation coefficient R2=0.77 with eight selected anthropometric parameters. Due to the high efficiency of prediction, the resulting knowledge was understandable and implementable so that it could be used and applied to a dataset of the Palestinian population.

References

1. Abdeen, Z., Jildeh, C., Dkeideek, S., Qasrawi, R., Ghannam, I., & Al Sabbah, H. (2012). Overweight and obesity among Palestinian adults: Analyses of the anthropometric data from the first national health and nutrition survey (1999-2000). *Journal of Obesity*, *2012*. https://doi.org/10.1155/2012/213547
2. Ahmed, F. E. (2005, August 6). Artificial neural networks for diagnosis and survival prediction in colon cancer. *Molecular Cancer*, Vol. 4, p. 29. https://doi.org/10.1186/1476-4598-4-29
3. AlAgha, A. S., Faris, H., Hammo, B. H., & Al-Zoubi, A. M. (2018). Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine. *Artificial Intelligence in Medicine*, *88*, 70–83. https://doi.org/10.1016/j.artmed.2018.04.009

4. Almadhoun, M. D., & El-Halees, A. M. (2017). Citation: Almadhoun MD, El-Halees AM (2017) Different Mining Techniques for Health Care Data Case Study of Urine Analysis Test. *Int J Biomed Data Min*, *6*(2), 129. https://doi.org/10.4172/2090-4924.1000129
5. Babajide, O., Hissam, T., Anna, P., Anatoliy, G., Astrup, A., Alfredo Martinez, J., … Sørensen, T. I. A. (2020). A machine learning approach to short-term body weight prediction in a dietary intervention program. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12140 LNCS*, 441–455. https://doi.org/10.1007/978-3-030-50423-6_33
6. Du, Y. C., & Stephanus, A. (2018). Levenberg-marquardt neural network algorithm for degree of arteriovenous fistula stenosis classification using a dual optical photoplethysmography sensor. *Sensors (Switzerland)*, *18*(7). https://doi.org/10.3390/s18072322
7. Ford, N. D., Patel, S. A., & Narayan, K. M. V. (2017). *Obesity in Low-and Middle-Income Countries: Burden, Drivers, and Emerging Challenges*. *7*, 36. https://doi.org/10.1146/annurev-publhealth
8. Gao, X., Xie, W., Wang, Z., Zhang, T., Chen, B., & Wang, P. (2020). Predicting human body composition using a modified adaptive genetic algorithm with a novel selection operator. *PLOS ONE*, *15*(7), e0235735. https://doi.org/10.1371/journal.pone.0235735
9. Jindal, K., Baliyan, N., & Rana, P. S. (2018). Obesity prediction using ensemble machine learning approaches. *Advances in Intelligent Systems and Computing*, *708*, 355–362. https://doi.org/10.1007/978-981-10-8636-6_37
10. Johnson, R. W., & College, C. (1996). Journal of Statistics Education, V4N1: Johnson. Retrieved September 7, 2020, from Journal of Statistics Education v.4, n.1 website: http://jse.amstat.org/v4n1/datasets.johnson.html
11. Hamdan, I; Awad, M; Sabbah, W. Short-Term Forecasting of Weather Conditions in Palestine Using Artificial Neural Networks. Journal of Theoretical & Applied Information Technology, 2018, 96.9.
12. Kupusinac, A., Stokić, E., & Doroslovački, R. (2014). Predicting body fat percentage based on gender, age and BMI by using artificial neural networks. *Computer Methods and Programs in Biomedicine*, *113*(2), 610–619. https://doi.org/10.1016/j.cmpb.2013.10.013
13. Kupusinac, A., Stokić, E., Sukić, E., Rankov, O., & Katić, A. (2017). What kind of Relationship is Between Body Mass Index and Body Fat Percentage? *Journal of Medical Systems*, *41*(1). https://doi.org/10.1007/s10916-016-0636-9

14. Luke, A., Durazo-Arvizu, R., Rotimi, C., Elaine Prewitt, T., Forrester, T., Wilks, R., … Cooper, R. S. (1997). Relation between body mass index and body fat in black population samples from Nigeria, Jamaica, and the United States. *American Journal of Epidemiology*, *145*(7), 620–628. https://doi.org/10.1093/oxfordjournals.aje.a009159
15. Awad, M; Zaid-alkelani, M. Prediction of Water Demand Using Artificial Neural Networks Models and Statistical Model. International Journal of Intelligent Systems and Applications, 2019, 11.9: 40.
16. Merrill, Z., Chambers, A., & Cham, R. (2020a). Development and validation of body fat prediction models in American adults. *Obesity Science and Practice*, *6*(2), 189–195. https://doi.org/10.1002/osp4.392
17. Merrill, Z., Chambers, A., & Cham, R. (2020b). Development and validation of body fat prediction models in American adults. *Obesity Science and Practice*, *6*(2), 189–195. https://doi.org/10.1002/osp4.392
18. Pereira, P. F., Serrano, H. M. S., Carvalho, G. Q., Lamounier, J. A., do Peluzio, M. C. G., do Franceschichichini, S. C. C., & Priore, S. E. (2010). Waist cicircumference as indicicator of body fat and metabolicic alterations in teenagers: Comparison among four references. *Revista Da Associacao Medica Brasileira*, *56*(6), 665–669. https://doi.org/10.1590/S0104-42302010000600014
19. Shahid, N., Rappon, T., & Berta, W. (2019, February 1). Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS ONE*, Vol. 14. https://doi.org/10.1371/journal.pone.0212356
20. Shao, Y. E. (2014). Body fat percentage prediction using intelligent hybrid approaches. *The Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/383910
21. SOCR Data BMI Regression - Socr. (n.d.). http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_BMI_Regression
22. Uçar, M. K., Uçar, Z., Köksal, F., & Daldal, N. (2021). Estimation of body fat percentage using hybrid machine learning algorithms. *Measurement: Journal of the International Measurement Confederation*, *167*, 108173. https://doi.org/10.1016/j.measurement.2020.108173
23. Walczak, S., & Cerpa, N. (2003). Artificial Neural Networks. In *Encyclopedia of Physical Science and Technology* (pp. 631–645). https://doi.org/10.1016/B0-12-227410-5/00837-1
24. WHO. (2015). *Palestine*. https://applications.emro.who.int/dsaf/EMROPUB_2016_EN_18926.pdf?ua=1

25. WHO. (2020). Obesity. Retrieved from WHO website: https://www.who.int/health-topics/obesity#tab=tab_1
26. Wang, M., & Chen, H. (2020). Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. Applied Soft Computing, 88, 105946.
27. Wang, Q. Q., Yu, S. C., Qi, X., Hu, Y. H., Zheng, W. J., Shi, J. X., & Yao, H. Y. (2019). Overview of logistic regression model analysis and application. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine], 53(9), 955-960.
28. Alabdallah, A; Awad, M. Using weighted support vector machine to address the imbalanced classes problem of intrusion detection system. KSII Transactions on Internet and Information Systems (TIIS), 2018, 12.10: 5143-5158.

التنبؤ بنسبة الدهون في الجسم على أساس القياسات الأنثروبومترية باستخدام نهج التنقيب عن البيانات

همسة عمرو، محمد عوض *

[1]قسم العلوم الصحية، المعلوماتية الصحية، الجامعة العربية الأمريكية، رام الله–فلسطين

h.amro@student.aapu.edu

[2] قسم هندسة علم الحاسوب ، الذكاء الاصطناعي ، كلية الهندسة وتكنولوجيا المعلومات ، الجامعة العربية الأمريكية– فلسطين

Mohammed.awad@gmail.com

ملخص

في السنوات الأخيرة، تم الإبلاغ عن أن أمراض القلب والسكري وبعض أنواع السرطانات هي أحد الأسباب الرئيسية للوفاة في معظم دول العالم. وتعد السمنة أحد أكثر عوامل الخطر شيوعًا لهذه الأمراض. وغالبًا ما تؤدي الدهون الزائدة في الجسم إلى السمنة. وقد تم تطبيق اكتشاف المعرفة من خلال النمذجة التنبؤية، وذلك للاستفادة من الكميات الهائلة من البيانات التي تنتجها أنظمة معلومات الرعاية الصحية. وتستخدم هذه الدراسة القياسات الأنثروبومترية بصفتها مدخلات لتقنيات استخراج البيانات المختلفة من أجل التنبؤ بنسبة الدهون في الجسم. وتم استخدام تقنية نقاط فيشر لاختيار الميزة الأكثر فعالية في عملية التنبؤ، والتي تتمثل بثمانية متغيرات هي: (البطن، ومؤشر كتلة الجسم، والصدر ، والورك، والعنق، والفخذ، والركبة، والعمر). وتم تطبيق ثلاثة مناهج للتنقيب عن البيانات؛ الانحدار المتعدد (MR)، وآلة ناقلات الدعم (SVM)، والشبكات العصبية الاصطناعية (ANN). وتفوقت ANN على الطرق الأخرى المستخدمة للتنبؤ بنسبة الدهون في الجسم من خلال معامل الارتباط R2= 0.77 مع ثمانية معلمات شملت القياسات البشرية والعمر. ونظرًا لتميز النمذجة بقدرتها العالية على التنبؤ، فقد كانت المعرفة الناتجة مفهومة وقابلة للتنفيذ، بحيث يمكن استخدامها وتطبيقها على مجموعة بيانات للسكان الفلسطينيين.

الكلمات الدالة: السمنة ، الارتباط ، التنقيب في البيانات ، الشبكة العصبية الاصطناعية ، التنبؤ.