Type of Training Recommendation Based on Body Fat Prediction Using LASSO Regression

1st Teddy Lioner
Faculty of Information Technology
Universitas Tarumanagara
Jakarta, Indonesia
teddy.535180014@stu.untar.ac.id

2nd Dyah Erny Herwindiati
Faculty of Information Technology
Universitas Tarumanagara
Jakarta, Indonesia
dyahh@fti.untar.ac.id

3rd Janson Hendryli
Faculty of Information Technology
Universitas Tarumanagara
Jakarta, Indonesia
jansonh@fti.untar.ac.id

2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA) | 978-1-6654-6136-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICITDA55840.2022.9971376

***Abstract*—Bodybuilding is a unique sport that has measurement of aesthetic, instead of performance. Body fat is one of significant metric in body building. Every bodybuilder, from beginner up to athlete level, is expected to have an ideal body fat according to their body preference, either bulky or slim body. Bodybuilder can create a more targeted exercise program by knowing their body fat percentage. However, measuring body fat percentage accurately tends to be expensive and cumbersome. Hence, the purpose of this research is to help every bodybuilder in calculating body fat percentage and giving tips about their current and future type of training easily and for free. The type of training recommendation is based on the predicted body fat and several answer for specific questions about their body preference. The model for predicting body fat is calculated using least absolute shrinkage and selection operator (LASSO) regression which achieved 73.43% in accuracy.**

Keywords—body fat, LASSO Regression, bodybuilder, body preference

I. Introduction

One of sustainable development goals is good health and well-being. To achieve good health bodybuilding is one of best sport to do. Bodybuilding is a unique sport. The uniqueness of bodybuilding is its focus on the beauty of the body, unlike other sports that prioritize performance. Many factors affect the beauty of the body, namely the shape of the body frame, muscle mass, and body fat percentage [1]. The shape of the skeleton is genetically inherited and muscle mass can be trained with resistance training. The body fat percentage of a bodybuilder must be considered in a certain range so that the beauty of the body can be seen clearly. Bodybuilder can create a more targeted exercise program by knowing their body fat percentage. However, measuring body fat percentage accurately tends to be expensive and cumbersome like DXA scan [2].

Predicting body fat is one of best way to recommend type of training for bodybuilder because there is specific range of ideal bodyfat to gain muscle and get toned body [3]. While it is possible to loss fat on some specific body parts [4] but it is entirely on the user's goal about what part of body to be worked on as different people has different type of aesthetic and different genetic aspect of muscle potential [5] on different region of body. So, the aim is to suggest the user for ideal bodyfat according to the user needs. To solve this problem, we propose an application to predict body fat percentage that is easy and cheap to do, so that bodybuilders can get references to design a better exercise program.

There is similar study predicting bodyfat targeted to women using Random Forest Regression, Extreme Gradient Boosting, Decision Tree, Support Vector Regression, Multilayer Perceptron Regression, and Least Square Support Vector Regression [2]. However, those regression method are not known to tackle the multicollinearity problem on bodyfat dataset. So in this study, Least absolute shrinkage and selection (LASSO) regression is a good choice for making a model prediction of body fat that probably has multicollinearity issue and readable formula to be used manually by human. LASSO regression has been used several times, such as in [6].

In this paper, we show the data taken from [7]. The actual data used for modelling in this study can't be disclosed because of confidentiality measurement. The data shown are only used to give example of dataset that the study used for LASSO regression to build a body fat prediction system.

The data have fifteen: density, body fat, age, weight, height, neck circumference, chest circumference, abdomen circumference, hip circumference, thigh circumference, knee circumference, ankle circumference, biceps circumference, forearm circumference, and waist circumference. Fourteen out of fifteen variables will be used for modeling. The one variable is not processed is density because the density data listed are used to calculate body fat as gold standard using SIRI equation, while the aimed user of application in this paper needs to avoid the trouble measuring density of body. So, independent factor data to measure body fat are derived from easier way such as chest circumference and the others.

II. Methods

A. LASSO Regression

LASSO, abbreviation of least absolute shrinkage and selection operator, is a linear regression variant that has a strategy shrinking the predictor variable regression coefficient. LASSO regression is well-suited for handling data with multicollinearity problem and elimination of redundant variable in a certain prediction model compared to standard linear regression [8]. LASSO is improved version of Ridge regression, then it is concluded LASSO is better to be used than Ridge Regression [6]. Ridge regression is able to tackle the multicollinearity prob but it will not eliminate the unnecessary variable by itself. LASSO is better cause it removes the unnecessary variable by itself. The implementation of LASSO comes in many ways such as LASSO with coordinate descent approach and the much recent one is LASSO with LARS approach (LASSO-LARS). The LASSO estimator is written as in Eq. 1 [8].

$$\hat{\beta}^{LASSO} = \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \beta_o - \sum_j^k \beta_j X_{ij}\right)^2\right\} + \lambda\sum_{j=1}^{p}|\beta j| \quad (1)$$

where Y denotes the dependent factor, X denotes the independent factor, β is a constant, n is the number of observations, p denotes the number of independent variables, and λ is the amount of shrinkage.

In this paper, the LASSO-LARS implementation is described into steps as shown [9]:

1. Find a proportional vector to the correlation vector among the error of predictor variables.

$$\hat{C} = X^T(y - \hat{\mu}) \quad (2)$$

2. Determine the greatest absolute correlation by using Eq. 3.

$$\hat{C} = \max\{|\hat{C}_j|\} \quad (3)$$

3. Determine X_A and the following A set as the active indices set that correspond to the predictor variable {1, 2, 3, 4, …, m} that is calculated by the value of largest absolute correlation as:

$$X_A = \{\dots s_j X_j^* \dots\}; j \in A \quad (4)$$

$$s_j = sign\{\hat{C}_j\}; j \in A \quad (5)$$

$$G_A = X_A' X_A \quad (6)$$

$$A_A = (X_A^T G_A^{-1} 1_A)^{-\frac{1}{2}} \quad (7)$$

4. Calculate the equiangular vector value. Equiangular vector can be defined as a vector dividing the X_A angle columns into equal size and have angle less than 90^o. The equiangular vector value is calculated using Eq. 8.

$$u_a = X_A \omega_A \text{ while } \omega_A = A_A G_A^{-1} 1_A \quad (8)$$

5. Determine the product vector

$$a = X' u_A \quad (9)$$

6. Calculate $\hat{\mu}_A$ with $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{y}\hat{\mu}_A$ to determine

$$\hat{y} = min_{j \in A^c}^{+}\{\frac{\hat{C}-\hat{C}_j}{A_A - a_j}, \frac{\hat{C}+\hat{C}_j}{A_A + a_j}\} \quad (10)$$

7. Repeat all steps for every variable selection until all predictor variables is selected.

At the end of steps repetition, the value $\hat{y}$ is calculated using formula

$$\hat{y} = \frac{\hat{C}_m}{A_m} \quad (11)$$

B. Multicolinearity

Multicollinearity is a condition in when there is a correlation between two or more predictor variables in multiple linear regression [9]. One of the metrics that can detect the presence of collinearity is the variance inflation factor (VIF). If the value of $VIF_{(j)} > 10$, then there is multicolinearity. The VIF value can be found using Eq. 12 [10]

$$VIF_j = \frac{1}{1-R_j^2}; j = 1,2,\dots,k \quad (12)$$

where R_j^2 represents the coefficient of determination of the predictor variable X_j.

C. Mean Square Error

Mean Squared Error or MSE is the average squared error between the actual value and the prediction value. The MSE is generally used to check the estimation of the error value in prediction model. A low MSE value or close to zero indicates that the forecasting results are in accordance with the actual data and can be used for prediction calculations in the future period [11].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (13)$$

D. Cross Validation

In LASSO regression, determining the best model is done by selecting the tuning parameter value that has the largest cross validation (CV) score. One method of cross validation is k-fold. To find the value of CV, Eq. 14 can be used [11]

$$CV = \frac{1}{2}\sum_{i=1}^{n} MSE_i \quad (14)$$

The cross-validation using k-fold produces k estimates of the MSE_1, MSE_2, …, MSE_k test errors. The cross-validation that should be used is 5-fold and 10-fold because it will produce validation values with high bias but low variance.

Some popular cross-validation modifications are GridSearch CV and BayesianSearchCV. GridSearchCV tries all combinations of values passed in the dictionary and evaluates the model for each combination using the cross-validation method. Therefore, after using this function we get the accuracy for each combination of hyperparameters and we can choose the one with the best performance.

BayesianSearchCV is derived from Bayesian optimization combined with the use of cross validation. Bayesian optimization is an approach to optimize objective functions that take a long time to evaluate. The principle is to build a surrogate function with a simpler form to approximate the actual objective function. The uncertainty of this surrogate function will be measured using Bayesian machine learning techniques, Gaussian process regression, and the acquisition function defined from this surrogate function to determine points that can be taken as samples.

III. Results and Discussion

A. Multicolinearity Analysis

Multicollinearity can be concluded by checking the VIF value of every feature that is available for making a model prediction. If the VIF value is more than 10, then it is most likely that the dataset has multicollinearity. The value of VIF in the dataset can be found on Table I.

TABLE I. VIF Value of The Dataset

Feature	VIF
Age	30.9430
Weight	250.5903
Height	404.1457
Neck	967.3165
Chest	1128.0578
Abdomen	879.5439
Hip	1850.1770

Thigh	1007.8823
Knee	1119.9708
Ankle	344.4232
Biceps	421.0373
Forearm	440.3816
Wrist	1246.1528

Every feature has VIF value more than 10. So, the dataset has multicollinearity problem. Despite that, LASSO regression is a fine regression method to make a model prediction out of this dataset.

B. Prediction Model Analysis

The ideal model accuracy based on R^2 value and CV score is ranged between 70%-80% since the dataset is not focused to measure specific group of people.

There is huge difference between R^2 and CV score. While CV score tells the whole dataset body fat prediction result after being randomized between 20-30% of test data several times and get average of it.

R^2 obtained from the degree of any linear correlation between actual body fat based on gold standard and predicted body fat based on the LASSO regression.

Finding the best tuning parameter for LASSO regression is to control the strength of the elimination, the stronger the elimination the more unnecessary variables are eliminated. The tuning parameter is ranged 0 to 1. This is conducted using these three methods from the Scikit-Learn library: LassoLarsCV, GridSearchCV, and BayesianSearchCV. The process of fitting each model gives the results as in Table 2. The LASSO-LARS regression method is used because of the VIF value that exceeded 10 on many variables which showed indications of multicollinearity.

In table II, the count of selected variable indicated how many variables are taken into features of LASSO regression model.

TABLE II. BEST TUNING PARAMETER

Metric	LassoLars CV	Grid SearchCV	Bayesian SearchCV
CV Score	0.7048	0.7047	0.7045
R2 Testing	0.6116	0.6117	0.6110
R2 Training	0.7343	0.7339	0.7354
MSE	20.8161	20.8081	20.8444
Tuning Parameter	0.89	0.91	0.82
Count of Selected Variables	5	5	5

In the LASSO regression model with CV-5 fold (LassoLarsCV), unsatisfactory results are obtained. Therefore, it is proposed to use modified cross validation such as GridSearchCV and BayesianSearchCV. It turns out that CV 5-fold, rather than CV modifications such as GridSearch and BayesianSearch, does not have much different results. Therefore, the LassoLars model with CV-5 fold as a predictive model is used. The model has R^2 accuracy of 73.43%, MSE value of 20.81, R^2 score from testing dataset of 61.16%, and cross validation score of 70.48%.

The Best Tunning Parameter (see Table II) in the training process shows that R2 has a value of around 73% for LassoLarsCV. This value is almost identical to the other two methods, Grid SearchCV and Bayesian SearchCV. R2 coefficient sometimes refers to as the "goodness of fit." A high R2 value indicates that the model fits the data well. LassoLarsCV has an R2 value of 73% indicates that 73 percent of the variation in the outcome has been explained by using the covariates included in the model. based on the model's performance, an R2 value that exceeds 70% can be said that the model is feasible.

Mean squared error (MSE) in regression analysis is often used as a measure of model evaluation. MSE is defined as the mean or Average of the square of the difference between actual and estimated values. MSE, in principle, measures the squared distance between model forecasting results and real observations. The smaller the mean squared error, the closer you are to finding the goodness of fit. MSE value of 20.8 means the average of square error sum within the prediction made, it can be made to root-MSE 4.56. Thus, we can infer that the prediction of bodyfat has room of error around 4.56. For example, if the predicted body fat is 25% that means the actual body fat according to the gold standard must be around 20.44% to 29.56 %.

To validate that the selection of the tuning parameter with CV-5 fold on the model is correct, manual checks are carried out through the following graphs.

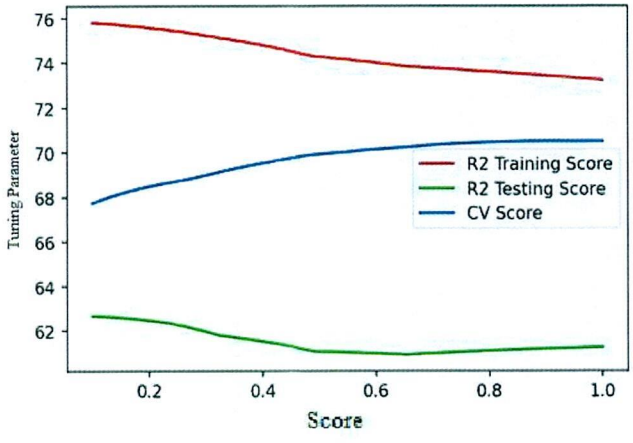

Fig. 1. Score Graph

Figure 1 shows a graph of R^2 and CV scores against the tuning parameter value. The CV score shows an increasing trend when the tuning parameter value approaches 1, which means the best tuning parameter value is the tuning parameter value close to 1. The R^2 score of the training dataset shows a decreasing trend as it approaches value of 1. This is reasonable because as it approaches 1, the selected variable taken to equation is lessened. Hence, the R^2 training score will lessen too. When the multicollinearity problem is overcome by LASSO regression, the R^2 value of the training dataset decreases while the most important CV score rises up almost making a convergence which indicates that the model is more reliable in handling data outside the training dataset.

The R^2 score of the test dataset shows a decreasing trend from the tuning parameter value of 0 to 0.7, which indicates

that the model is not yet stable. This is confirmed through Figure 2 that shows the MSE plot. The MSE value shows an upward trend giving the same meaning as the R^2 score of the test dataset. From the graph of the MSE and R2 values of the test dataset, it shows that the LASSO regression model is stable in the tuning parameter value range of 0.7 to 1 with the trend of the R^2 value of the test dataset increasing and MSE decreasing.

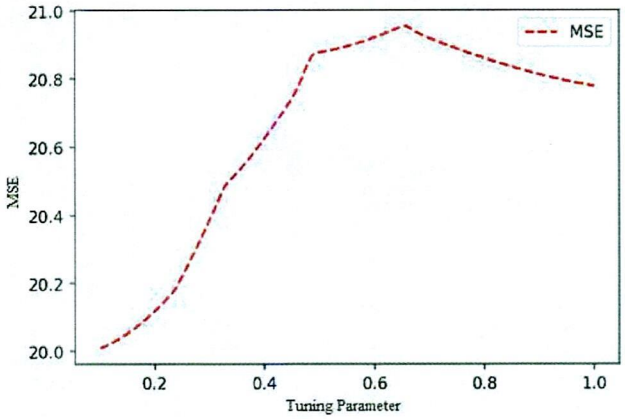

Fig. 2. MSE Graph

The MSE value with tuning parameter approaching 0 is indeed the lowest because MSE compares the bias value with the full model least squares coefficient estimator. Therefore, the MSE value becomes a benchmark to see the best tuning parameter value. There is a decrease in the tuning parameter value after the MSE value increases from the tuning parameter value that is equal to 0.

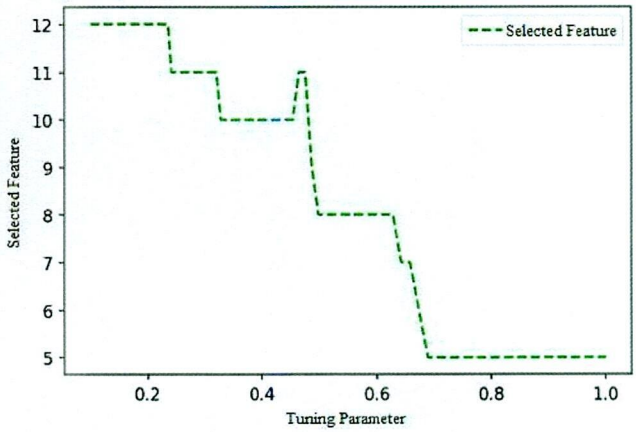

Fig. 3. Number of Selected Feature Graph

LASSO regression is a regression that provides a variable selection effect so that the variables used in the regression model is less than the independent variables used as input data. Figure 3 shows a graph of the number of features, selected variables, in the model. The minimum number of variables for the regression model obtained is 5 features with a tuning parameter value within range of 0.7 to 1.

From the analysis of the metric values above, it shows that the best tuning parameter value is within range of 0.7 to 1. CV-5 fold gives a tuning parameter value of 0.89 which is in the range of 0.7 to 1, so the conclusion is that a tuning parameter value of 0.89 can be validated as the best tuning parameter value. The formula for calculating body fat can be written as Eq. 15.

$$body\ fat = -0.2102 * weight\ (kg) - 0.0561 * height(cm) - 0.07534 * neck(cm) + 0.9024 * abdomen(cm) - 0.0643 * hip(cm) - 28.3397 \quad (15)$$

Based on body fat prediction equation Eq. 15, the most significant body part affecting is abdomen as it has highest coefficient value. While the most insignificant body part affecting body fat is height. It is in accordance to [12] that shows abdomen area has most correlation to body fat and height has the most insignificant correlation.

C. *Training Recommendation Type Based on Body Fat*

The process to determine training recommendation type is shown in the Figure 4.

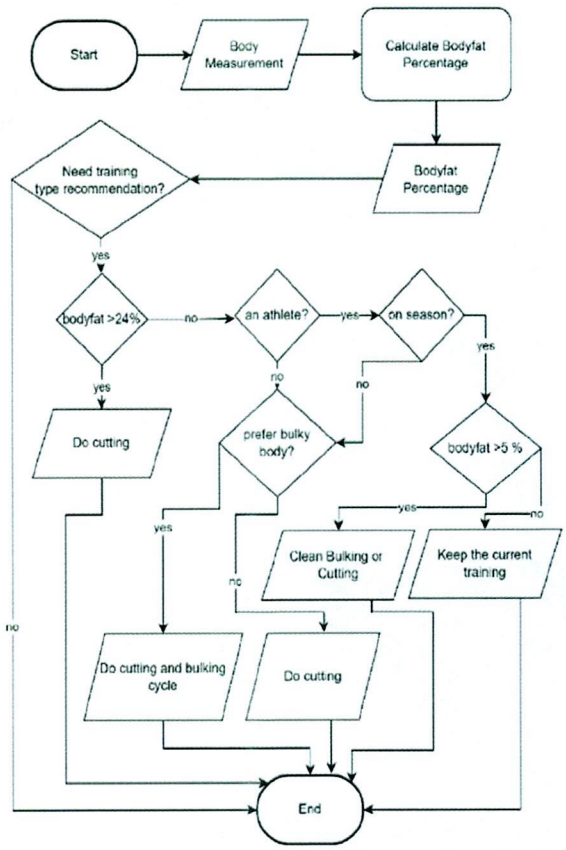

Fig.4. Flowchart of System

Regarding athlete and non-athlete recommendations, it differs when a bodybuilding athlete is preparing for competition (on the season), in contrast to bodybuilders who try to limit body fat to the threshold of 5% or below [1, 13, 14]

In general, there are four training type recommendations as shown below:

1. Do Cutting

   This training type is recommended for people who are obese. Otherwise, someone who has a healthy body fat percentage and body preference as not as bulky will do cutting to keep their body in shape and avoid

getting too big. Cutting is training that maintains calorie intake below a bit of daily calorie need.

2. Do Cutting and Bulking Cycle

   This training type is recommended for someone who prefers a bulky body and gets big as he can get. Cutting will be done to achieve a healthy body fat percentage, 24% and below. Doing bulk will be done after someone has a healthy body fat percentage and get muscle mass as fast as he can. Bulking is training that maintains calorie intake above a bit of daily calorie need.

3. Do Clean Bulking or Cutting

   This training type is recommended for a bodybuilder who is preparing for competition. Body fat percentage should be below or equal to 5%. Such low body fat helps bodybuilders to achieve peak aesthetic and a greater chance of winning. So, the bodybuilder either loses fat or gains more muscle without adding more fat. Losing fat can be done through Cutting. Gaining muscle without adding much more fat can be done by clean bulking. Clean Bulking is training that maintains calorie intake as much as daily calorie need. Clean bulking is much harder to do than typical bulking.

4. Keep the Current Training

   This training type is recommended for a bodybuilder who prepares for competition and has achieved a body fat percentage that is lower or equal to 5%.

In the system, users input their body measurements and calculate the body fat percentage as the output. There are additional inputs for those who want to get a recommendation for training type in the form of several questions. The output from these additional inputs is a recommendation for the type of training. The system is available at https://web-lemak-skripsi.herokuapp.com. Below is provided an example of bodyfat calculation and recommendation.

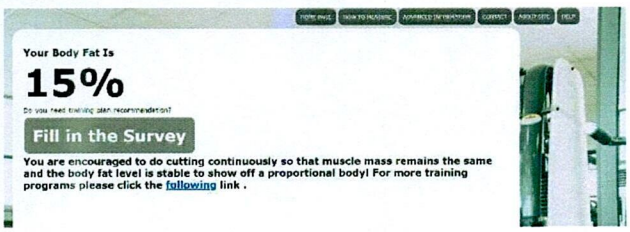

Fig.5. Example of System Result Do Cutting

D. System Usability Scale Analysis

The system is measured by System Usability Scale (SUS) score. System Usability Scale is one of the best time-proven among usability test conducted for decades, so this study will use the SUS score to analyze the system. System Usability Scale is served as a questionnaire that has 10 statements that need to be asked if the respondent agrees with them. The SUS score has a range of 1-100. The score is better if the score is higher. For further technical details of SUS can be accessed on [15].

There are two main groups to evaluate the system. The first group is represented by one personal trainer as someone who is expert in bodybuilding field. The second group is represented by one layman who just started his bodybuilding history less than 1 year to show the casual user in bodybuilding field.

The first respondent is a personal trainer and expert on bodybuilding. This kind of user is most likely to use the website more than average people. The first respondent gives a SUS score of 95. The second respondent represents a casual bodybuilder in bodybuilding. This respondent gives a score of 82.5. The result is considered as excellent.

The SUS score shows a personal trainer has a higher SUS score on using the system than new bodybuilder. This is expected because personal trainer is expected to do more thorough research, especially bodyfat, to build a training plan than a casual bodybuilder. Even so, the system achieves great result too with casual bodybuilder with a score of 82.5.

IV. Conclusion

In this paper, we explore the usage of the LASSO regression to predict body fat percentage and recommend the type of training suitable for the user. Generally, this system can be used by beginners who want to exercise and be healthy and also professional athletes or bodybuilders who want to control their body fat percentage precisely and design their training regime. The user of this system can use a cheap body scale and a measuring tape to get weight (kg), height (cm), the circumference of the neck, abdomen, and hip (cm), then input them into the system. After the input is done, the user will get a body fat percentage shown. The user also can get a recommendation for training type after several questions provided are answered.

LASSO proved its capability to handle multicollinearity in a dataset. Several Cross-Validation is done to find the best tuning parameter value of the LASSO regression model, which is 0.89. The LASSO regression model predicts body fat percentage with a coefficient of determination, R2, of 73.43% and MSE of 20.8161.

Even so, the study is limited by sample size and profile available when data collection was held. Future work can be improved by adding specific dataset for each group of people casual, athlete, gender, and so on. So, the accuracy of model can be improved.

References

[1] Barben, "How bodybuilding is judged", https://barbend.com/how-bodybuilding-is-judged/, (accessed September 7, 2021).

[2] S. S. A. Alves et al., "Gender-based approach to estimate the human body fat percentage using Machine Learning," 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533512.

[3] E. R. Helms, A. A. Aragon, and P. J. Fitschen, "Evidence-based recommendations for natural bodybuilding contest preparation: Nutrition and Supplementation," *Journal of the International Society of Sports Nutrition*, vol. 11, no. 1, 2014.

[4] A. Paoli et al., "effect of an endurance and strength mixed circuit training on regional fat thickness: the quest for the 'spot reduction," International Journal of Environmental Research and Public Health, vol. 18, no. 7, p. 3845, Apr. 2021, doi: 10.3390/ijerph18073845.

[5] S. M. Roth, "Genetic aspects of skeletal muscle strength and mass with relevance to sarcopenia," *BoneKEy Reports*, vol. 1, 2012, doi:10.1038/bonekey.2012.58

[6] D. Budi, D. E. Herwindiati and J. Hendryli, "Land use change using least absolute shrinkage and selection operator regression in jakarta's buffer cities," 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2021, pp. 30-35, doi: 10.1109/ISCAIE51753.2021.9431770.

[7] K. W. Penrose, A. G. Nelson, and A. G. Fisher, "Generalized body composition prediction equation for men using simple measurement techniques." 1985. doi: 10.1249/00005768-198504000-00037.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R. Stat. Soc. Ser. B, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517- 6161.1996.tb02080.x.

[9] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani "Least angle regression," The Annals of Statistics, Ann. Statist. vol. 32, no. 2, pp. 407-499, April, 2004. doi: 10.1214/009053604000000067

[10] J. Snell, D. C. Montgomery, and G. C. Runger, "Applied statistics and probability for engineers." 1995. doi: 10.2307/2983314.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An introduction to statistical learning: with applications in R." 2021. doi: 10.1007/978-1-0716-1418-1.

[12] Albulescu, Dana, and Adriana Iliescu, "Correlations between areas, volumes or body fat and anthropometric variables." Current health sciences journal vol. 40, no. 2, 2014. pp. 116-118. doi:10.12865/CHSJ.40.02.06

[13] Liz, "The differences between the bulking and cutting phases," https://www.nestacertified.com/the-differences-between-the-bulking-and-cutting-phases/. (accessed October 13, 2021).

[14] Ace Fitness, "Guideline for body fat loss," https://www.acefitness.org/education-and-resources/lifestyle/blog/112/what-are-the-guidelines-for-Percentage-of-body-fat-loss/, (accessed September 7, 2021).

[15] J. Brooke. "SUS: A quick and dirty usability scale". Usability Eval. Ind. 189, 1995, doi: 10.1201/9781498710411-24