Body fat percentage calculation: A Linear Regression Model on Croatian tennis players morphology

Matea Vasilj*, Vlatko Vučetić‡, Marko Sukreški‡, Dario Bojanjac*
* Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
† Faculty of Kinesiology, University of Zagreb, Croatia
‡ Ultrax technologies, Zagreb, Croatia
matea.vasilj@fer.hr

***Abstract*—Body composition is an important aspect of the fitness and performance of tennis players. Measuring body composition accurately can be a challenging task due to its complex nature and various factors that can affect the results. This study aims to examine the relationship between tennis player morphology and body fat through mathematical models, reducing the reliance on costly measurement equipment. Morphology of Croatian tennis players dataset was used, and the results of statistical analysis showed a strong correlation between player morphology, age, sex and body fat percentage. Furthermore, a PCA analysis was carried out which revealed that male and female populations become distinct when they hit puberty and allows us to explain the differences in body composition between the sexes. Using these results, a set of linear regression models were developed that can be used to predict body fat percentage. This research adds to the existing literature by highlights the significance of factors affecting body fat percentage in tennis players, which are skin fold thickness, age, and sex. The newly developed models offer a cost-effective solution for coaches to assess and monitor body composition in tennis players with skin fold measurements instead of more expensive equipment.**

Keywords*—body fat percentage, tennis player morphology, tennis, data analysis, machine learning

I. INTRODUCTION

Body composition has long been of interest to researchers in the fields of kinesiology and sports science. It is an important factor in determining athletes performance, and it is widely recognized that higher levels of body fat are associated with lower levels of performance in various sports. The Faculty of Kinesiology has been performing periodic measurements on athletes since the early 2000s, and was able to provide us with the Croatian tennis player morphology data needed for this research. In tennis, body composition is especially important as it can affect factors such as speed, agility, and endurance. However, measuring body composition accurately and reliably can be challenging due to the cost and practical limitations of traditional methods such as dual-energy x-ray absorptiometry (DXA) and underwater weighing, especially for large groups of athletes [1].

In this study, we aimed to develop a more cost-effective and practical solution for measuring body composition in tennis players. We used a dataset containing morphology data of tennis players, and applied statistical analysis, principal component analysis (PCA), and linear regression to identify the most important factors that influence body fat percentages. The results of our analysis allowed us to develop formulas for estimating body fat levels in total and for specific body parts. Furthermore, based on the same data, the critical age at which boys and girls start to differ in their morphological profile was also identified in order to help coaches in planning and modeling the training process.

This study is significant as it provides a practical and cost-effective solution for measuring body composition in tennis players. It also provides new insights into the factors that influence body fat levels in this population. By providing coaches and trainers with this information, we hope to improve understanding of body composition in tennis players and to inform future research in this area and hopefully transfer them to other sports as well.

II. RELATED WORK

Body composition is a well-studied topic in the field of sports science, and there is still a growing body of research on this topic. Previous studies have investigated body composition in a variety of sports, including wrestling [2], soccer [3], and basketball [4]. These studies have shown that body composition is an important factor in determining athletic performance, and that higher levels of body fat are associated with lower levels of performance.

A number of methods have been proposed to estimate body fat, including skin fold thickness measurements, circumference measurements, and DXA but each of these methods have their own limitations. One of the biggest problems is the requirement of costly equipment and time, which can be an obstacle for sports organizations and athletes. DXA is considered the gold standard for body composition assessment, but it is also expensive and exposes the athletes to low levels of ionizing radiation. Moreover, the Tanita scale is a widely used device for body composition analysis and measurement of body fat percentage. It uses bioelectrical impedance analysis (BIA) technology to estimate body fat percentage by sending a low-level electrical current through the body and measuring the resistance [5]. On the other hand, skin fold thickness measurements have seen widespread utilization as a cost-effective and convenient method to estimate body fat. References [6], [7] provide evidence for the validity and reliability of skin fold thickness measurements as a method for estimating body fat, and highlight the simplicity and low cost of this method compared to other body fat assessment techniques.

There is also a growing interest in using statistical methods to analyze body composition data. For example, some studies have used regression analysis to predict body fat percentages based on skin fold thickness and circumference measurements, respectively [8], [9]. PCA was used to identify the most important factors that influence body fat levels [10]. However, most of these studies have focused on specific sports or age groups, and there is a lack of information about body composition in tennis players.

This study aimed to estimate total and regional body fat percentages using skin fold thickness, age, and gender in regression analysis. Our contribution is a comprehensive analysis of body composition in tennis players, identifying key factors that impact body fat. The insights gained have potential to guide future research in the field.

III. Dataset description

A. Structure and Types of Data

The tennis player morphology measurements were collected over a period of more than twenty years at the Faculty of Kinesiology's diagnostic center. The dataset includes measurements from 828 athletes, 370 of them being female and 458 being male. The athletes range in age from 7 to 57 years old and come from different sports clubs and have undergone multiple measurements over the years. It's worth mentioning that 718 of the participants are less than 18 years old and that the majority of the measurements were taken as part of the athletes' schooling requirements. This provides a unique opportunity to analyze the development of body composition and physical performance in young athletes.

The data used in the study consists of 73 features, including age, sex, and various morphology data. These features can be grouped into different measurement types, such as widths, lengths, circumferences, diameters, and most importantly, skin fold measurements. Additionally, the data includes body fat percentage measurements collected using the Tanita scale. Skin fold thickness measurements were taken using calipers at specific anatomical sites on the body, including the tricep, bicep, subscapular, suprailiac, abdominal, thigh, and calf, to estimate body fat percentage. It is important to mention that all used features are on the same scale. This helps to ensure that all the variables have equal importance and contribution in the regression models, and it also helps to mitigate any issues related to numerical stability or computational efficiency during the optimization process. Additionally, having all the features on the same scale allows for easy comparison and interpretation of the results, and it ensures that the variables do not dominate each other, leading to more robust and accurate predictions.

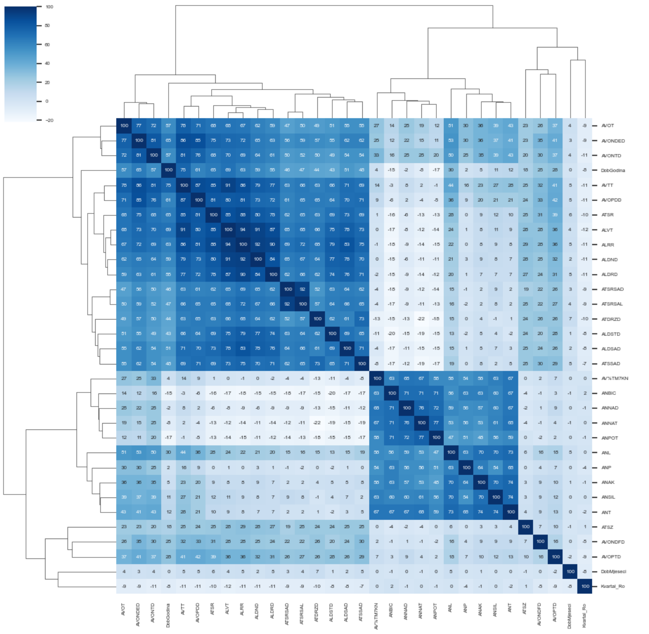


Fig. 1: Clustermap of Data Correlations

B. Data scarcity and high correlations

Having such a high dimensional dataset, we noticed a high degree of correlation among many of the features. A visual representation of this was seen through the cluster map shown in Fig. 1, which showed a clear grouping of features into two consecutive groups. This high degree of correlation is not uncommon in large datasets and can make interpretation of the data more challenging. To address this issue and improve the understanding of our data, we employed dimensionality reduction techniques. These techniques help to reduce the high-dimensional data into a lower-dimensional space, allowing for more clear and meaningful analysis. The use of dimensionality reduction will enable us to better understand the relationships between the features and how they impact the overall performance of the athletes.

Another challenge regarding the dataset are the missing values. Considering that the measurements were performed by different experts over many years, there was no guarantee that everyone would perform all the measurements found in our feature set. To handle this problem, we decided to replace the missing values with the column means if more than half of the data in that column was non-empty. If a column had less than half of the data filled, it was discarded.

IV. Methods

A. PCA

One of the research questions was related to identifying the differences between male and female players. The

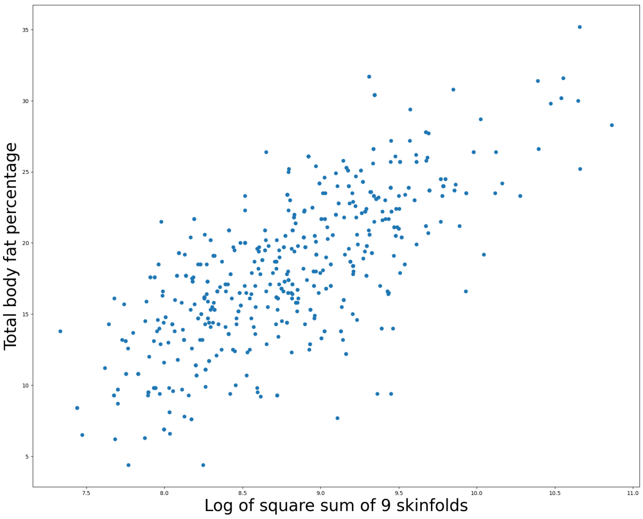


Fig. 2: Relationship between skin folds and body fat percentage

first step was to conduct a PCA analysis on all athletes using their sexes as targets. Furthermore, considering the heterogeneity of the age groups, we decided to additionally perform a PCA analysis over each age group.

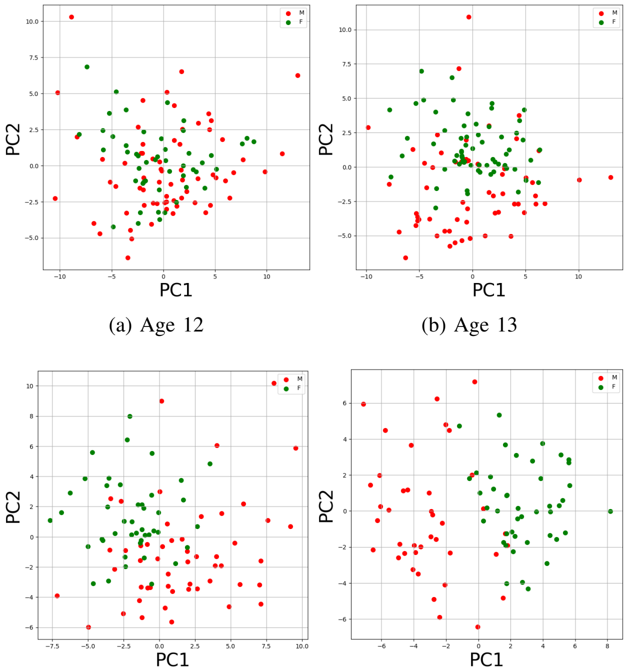
The knee analysis was conducted to determine the number of components necessary to cover the most variance in each age group. The results showed that the first two components alone covered around 60 percent of the variance for each age group and additional components had a minimal contribution in variance explanation.

B. Linear regression

Linear regression was used to determine the ability to calculate the total body fat percentage and of each body part using the Tanita scale measurements. The implementation of the linear regression was performed using the statsmodels library. The primary aim of this analysis was to examine the relationship between the Tanita scale measured body fat percentages and determine whether a prediction model could be developed.

In order to determine the best set of independent features, feature selection methods from the scikit-learn library were used to identify the most important predictors for body fat percentage calculation. Specifically, the SelectKBest, Recursive Feature Elimination (RFE) and SelectFromModel methods were applied. They were used to determine which features make the biggest impact in the linear regression model for body fat percentage prediction. Number of features as a hyperparameter ranged from 2 to 10.

After observing that most of feature selection methods choose skin fold thickness measurements along side with age and sex as the most important features, we decided to take a closer look at the impact of each skin fold and its relationship to the target variable and relationship between each other. Given the still present high correlations between skin fold thickness measurements, and taking into consideration the findings from Jackson and Pollock's research, we decided to aggregate the skinfold measurements into a single feature for our linear regression model [11].



(a) Age 12

(b) Age 13

(c) Age 14

(d) Age 15

Fig. 3: PCA Analysis of Age-Related Differences between Boys and Girls

Initially, we observed a logarithmic relationship between the sum of the nine skin folds and body fat percentage. After performing a log transformation on the sum, we noticed a slight quadratic relationship. However, after squaring the log of the sum, we finally obtained a linear relationship. Linear relationship between mentioned final feature and body fat percentage in total can be seen in the Fig. 2. After repeating the research for other body fat percentages, we obtained similar results. Final formula for combined feature for body part is shown in (1).

$$cmb\_ft = \log \left(\sum_{i=1}^{9} skinfold_i \right)^2 \tag{1}$$

V. Results

A. PCA

The PCA analysis revealed that female and male players start to differ at the age of 13. This is particularly important as the majority of the players in the dataset were below the age of 18. The PCA plots in Fig. 3 clearly showed the age-related differences between the two groups, highlighting the need for further research to understand these differences in greater detail.

B. Linear regression

The comprehensive search for the optimal hyperparameters was conducted by trying various combinations and evaluation of the linear regression models results. Evaluation was performed using measures from statsmodels, including adjusted R-squared, F-statistic and p-values among others, to assess the goodness of fit and statistical significance of the model.

The F-statistic is a measure of the overall significance of the regression model. It is calculated as the ratio of the explained variance to the residual variance in the model, and it is used to determine if the regression coefficients are significantly different from zero. The F-statistic is used in combination with the p-value to determine whether the model is statistically significant. The adjusted R-squared indicates how well the independent variables in a regression model predict the dependent variable, taking into account the number of independent variables in the model. It adjusts the R-squared value by penalizing the model for the addition of unnecessary variables, allowing for a better comparison of models with different numbers of predictors.

Before presenting the resulting formulas, we introduce binary indicator variable for gender, M, with a value of 1 indicating that the person is male and a value of 0 indicating that the person is female.

The results of the best linear regression models used to predict body fat percentage are summarized in Table I. The table provides the formulas and the adjusted R-squared values for each body part: Whole body, Abdominal, Leg, and Hand.

All four of our final models have high F-statistic indicating that the model as a whole is a good fit, and that the independent variables are important predictors of the dependent variable. It can be used to compare the fit of different regression models, with higher F-statistics indicating a better overall fit. Furthermore, low p-value (less than 0.05) for our predictors suggests that they are significantly related to the response and provide strong evidence against the null hypothesis that the predictor is not related to the response.

The factor of "M" (indicator for male gender) also plays a crucial role in all provided formulas, as it modifies the intercept term. This finding is consistent with the established knowledge that males generally have a lower body fat percentage [12].

The formulas developed in this study can be compared to the widely used Jackson and Pollock equations for estimating body fat percentage. These equations have been validated and widely used in the field, but have limitations as they use a small set of body circumferences and skin fold thickness measurements, and have limitations in their accuracy for certain populations and they only estimate total body density. On the other hand, the formulas developed in this study utilize a larger set of data and a linear regression approach, potentially leading to more accurate results for the specific population and data set used. A statistical test was performed in order to determine the difference between Jackson and Pollock's equations and our model. A random sample of 5 entries was taken from the models and a t-test was conducted. The results showed a significant difference between the models, with a p-value of 0.024, indicating that the models produce significantly different outputs. On the other hand, when comparing the Jackson and Pollock's results with actual measurements, a significant difference was found at a p-value of 0.0001. However, it is important to note that the validity and accuracy of the formulas developed in this study should also be thoroughly tested and validated in independent samples before widespread use.

It is commonly accepted that age is positively correlated with body fat percentage. As people age, their body fat percentage tends to increase, which is often attributed to a decline in muscle mass, hormonal changes, and a decrease in physical activity. This positive correlation between age and body fat percentage is reflected in our linear regression models, where age is consistently associated with higher body fat percentage across all four body parts. The results are in line with previous studies that have found similar trends.

Finally, it is reasonable to expect that body fat percentage would have a positive correlation with the tailored feature representing the sum of all skin folds, as the skin fold thickness is a commonly used measure of subcutaneous fat. The more body fat a person has, the thicker the skin fold measurement will be. This relationship between skin fold thickness and body fat percentage has been well established in the scientific literature, and the use of skin fold thickness as a predictor of body fat percentage is widely accepted in the field of human anatomy and physiology.

VI. Discussion

The results of our study provide valuable new insights into body composition in tennis players. Our findings suggest that body fat levels are influenced by a number of factors, including age, sex, and skin folds, and that these factors differ between boys and girls. By developing model for measuring body fat levels in total and for specific body parts, we have provided a practical and cost-effective solution for measuring body composition in tennis players.

Our study has some limitations that should be noted. The sample size of our study was relatively small, and our findings may not be representative of all tennis players. Additionally, our data were collected using a single measurement technique (the Tanita scale), and it is possible that other measurement techniques would yield different results.

Despite these limitations, our study provides valuable new information about body composition in tennis players, and it has the potential to inform future research in this area. By providing coaches and trainers with practical and cost-effective solutions for measuring body composition,

TABLE I: Body fat percentage calculation formulas

Body part	Formula	Adjusted R-squared
Whole body	$y_1 = cmb\_ft * 0.1634 + age * 0.3742 - 2.563 * M$	95.6%
Abdominal	$y_2 = cmb\_ft * 0.1180 + age * 0.3352 - 1.694 * M$	94.5%
Leg	$y_3 = cmb\_ft * 0.2061 + age * 0.5429 - 4.256 * M$	94.6%
Hand	$y_4 = cmb\_ft * 0.2117 + age * 0.5251 - 4.1068 * M$	94.7%

we hope to improve our understanding of this important factor in athletic performance.

Future studies could build upon our findings by collecting data from larger and more diverse groups of tennis players, and by using different measurement techniques to validate our results. Additionally, it would be interesting to investigate the relationships between body composition and other physical and performance measures, such as speed, agility, and endurance, in tennis players. These studies could provide additional insights into the factors that influence athletic performance in this population. Additionally, expanding the data collection to include various sports could lead to the discovery of generalizations across disciplines. This information could be utilized to develop formulas that can be applied to multiple sports.

VII. Conclusion

Our study provides valuable new insights into body composition in tennis players. We have developed practical and cost-effective solutions for estimating body fat levels in total and for specific body parts, using formulas provided in the text. We compared our models to the Jackson and Pollock's, widely considered state of the art, to evaluate our results. Our findings showed significant differences not only between our estimations and theirs, but also between their estimations and obtained measurements. The most important factors that influence body fat percentages were identified. We provided a comprehensive analysis of body composition in tennis players and reveals that boys and girls start differing in their morphology traits at the age of thirteen, which can inform training modeling for coaches.

This research has the potential to shape future studies on body composition in tennis players and to deepen our understanding in this area. Providing affordable and efficient models for determining body composition, this work aims to support the development and improvement of athletic performance in tennis players and inform coaches and trainers.

Acknowledgment

The authors would like to express their gratitude to the Faculty of Kinesiology's diagnostic center for providing the data used in this study and to Ultrax technologies for their expertise in the field of kinesiology, which was invaluable to the successful completion of the research.

References

[1] R. B. Mazess, H. S. Barden, J. P. Bisek, and J. Hanson, "Dual-energy x-ray absorptiometry for total-body and regional bone-mineral and soft-tissue composition," *The American journal of clinical nutrition*, vol. 51, no. 6, pp. 1106–1112, 1990.

[2] E. Franchini, C. J. Brito, and G. G. Artioli, "Weight loss in combat sports: physiological, psychological and performance effects," *Journal of the international society of sports nutrition*, vol. 9, no. 1, p. 52, 2012.

[3] P. Aagaard, J. L. Andersen, and P. Dyhre-Poulsen, "Training-related changes in body composition and muscle strength in young soccer players," *Journal of Strength and Conditioning Research*, vol. 24, no. 12, pp. 3268–3275, 2010.

[4] B. Vincent, G. Tolhurst, Z. Crowley-McHattan, and M. Batterham, "Body composition and physical performance of adolescent basketball players," *Journal of Strength and Conditioning Research*, vol. 24, no. 7, pp. 1824–1831, 2010.

[5] D. L. Simmons, M. C. Gonzalez, C. Bouchard, E. Ravussin, S. Lillioja, W. C. Knowler, P. A. Tataranni, and A. D. Salbe, "Validity of bioelectrical impedance analysis and skinfold thickness measurements for estimation of percent body fat," vol. 66, pp. 2085–2089. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15239930

[6] F. Katch, W. McArdle, and V. Katch, "Body composition and physical performance," *Medicine and Science in Sports*, vol. 13, no. 1, pp. 15–19, 1981.

[7] T. G. Lohman, A. F. Roche, and R. Martorell, "Prediction of body density of males from skinfolds and circumferences," *Human biology*, vol. 60, no. 2, pp. 709–723, 1988.

[8] F. I. Katch, W. D. McArdle, and V. L. Katch, "Body composition and physical performance," *Medicine and Science in Sports and Exercise*, vol. 13, no. 1, pp. 15–19, 1981.

[9] T. G. Lohman, M. H. Slaughter, R. A. Boileau, J. H. Himes, A. P. Rocchini, and S. Weller, "Prediction of body density of males from skinfolds and circumferences," *Human Biology*, vol. 60, no. 2, pp. 709–723, 1988.

[10] C. Bouchard, A. Tremblay, J. P. Després, A. Nadeau, P. J. Lupien, G. Theriault, J. Dussault, S. Moorjani, S. Pinault, and G. Fournier, "Genetic and nongenetic determinants of regional fat distribution," *Metabolism*, vol. 39, no. 3, pp. 324–329, 1990.

[11] A. S. Jackson and M. L. Pollock, "Generalized equations for predicting body density of men," *British Journal of Nutrition*, vol. 40, no. 03, pp. 497–504, 1978.

[12] A. Furnham and K. Cheng, "Gender differences in body fatness and fat distribution," *International Journal of Obesity*, vol. 26, no. 3, pp. 381–390, 2002.