Estimation of Anthropometric Measurements Using Optimized Machine Learning Models with Bayesian Algorithm

Didem Guleryuz[1], Ömer Faruk Efe[2], Burak Efe[3*]
[1]Department of Management Information Systems, Bayburt University, Bayburt, Turkey, e-mail: guleryuz8687@gmail.com
[2]Department of Industrial Engineering, Bursa Technical University, Bursa, Turkey, e-mail: omerfarukefe86@gmail.com
[3]Department of Industrial Engineering, Necmettin Erbakan University, Konya, Turkey, e- mail: burakefe0642@gmail.com, Mobile number: +905059601869

Corresponding Author:
Burak Efe
Email: burakefe0642@gmail.com

Acknowledgments

This work was funded by Necmettin Erbakan University under scientific research project code 221219008.

Abstract

This study collects the anthropometric measurements and weights of 185 male individuals between 55 and 65 years old from Ankara city of Turkey. A total of 29 variables with three inputs and twenty-six outputs are collected. This paper aims to develop machine learning-based models to estimate anthropometric measurements from weight, stature, and eye height. These models are support vector regression (SVR) optimized with Bayesian based on quadratic kernel, Gaussian Process Regression (GPR) optimized with Bayesian based on matern5/2 kernel. This study contributes to SVR and GPR models by using Bayesian method to optimize the parameters as a difference from the literature. According to the literature review, applying these two models to anthropometric measurements for the first time is predicted. The estimation results are compared based on three metrics, namely Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE). GPR optimized with Bayesian model has better accuracy than SVR optimized with Bayesian for all combinations except interpupillary distance, according to the obtained results. The RMSE values of the best models selected for each combination varied between 0.255 and 0.319 during the testing phase. Especially the estimations made with GPR optimized with Bayesian have a shallow error rate.

Keywords: Anthropometric measurements, machine learning, estimation models, parameter optimization, support vector regression, Gaussian process regression

1. Introduction

Anthropometric data is used to make designs (intersection for human-environment), considering the size differences of users of all tools and equipment (which vary according to age and gender). Height and weight measurements are important indicators in the follow-up of growth, the detection of nutritional diseases, the fields of energy consumption, and patients' health care. These two parameters may not always enable reliable estimation and measurement [1]. Height and weight can play an effective role in determining other human anthropometric data. In particular, the body measurements dealt with first in health-related issues are height and weight. W. C. Chumlea was one of the first researchers to propose linear equations for estimating height and weight from anthropometric measurements for the elderly population. Different studies have also been conducted to estimate weight and height. Linear equations are quite interesting. Because its representations are easy, understandable and have an analytical solution. It can be easily applied by a specialist [2-3].

Concurrent with Chumlea's work, the field of Machine Learning (ML) and statistical and probability theory models; research, engineering, economics, health, etc. started to play an important role in the field [4-6]. ML is closely related to computational statistics and is defined as developing algorithms that learn and make predictions from data or experience. ML algorithms can find patterns that are often impossible in complex scenarios for humans to identify. Therefore, ML algorithms can give more accurate results than regression models [7].

Recently, kernel machines have been presented as a suitable approach for regression of biometric data. As noted by Scholkopf, kernel machines provide modularity in design, allowing for easy combination. Networks to be tuned with different learning algorithms and compared to other models such as neural with few parameters ensures minimal in a pseudo-local optimization procedure [8-9].

Chumlea and Guo (1992) presented a linear equation to determine the body stature using knee height [10]. Michels et al. (1998) handled body height and weight from anthropometric measurements [11]. Kaya et al. (2003) introduced adaptive neuro-fuzzy inference system to estimate anthropometric measurements as an alternative to stepwise regression analysis [12]. Gauld et al. (2004) used a linear regression to calculate the height based on ulna length and age [13]. Hu et al. (2007) examined 47 anthropometric dimensions and three items of functional strength [14]. Kuiti and Bose (2016) developed predictive equations for height estimation using knee height in elderly nutrition. Multiple regression analyses were performed to generate stature predictive equations using age, weight and knee height as independent variables [15]. Lee et al. (2018) searched the effects of cold and heat patterns to the anthropometric measures for men and women individuals. Firstly, they used wrapper-based variable selection technique to define to be examined antropometric measures. Then, they handled Naïve Bayes and logistic regression methods to examine the relationships between them. They found that the most important indicators are body mass index and rib circumference in women and body mass index in men [16]. Ferenci ve Kovacs (2018) examined how well can body fat percentage be predicted from easily measureable data such as age, gen-der, weight, height, waist circumference and different laboratory results. They applied linear regression, feedforward neural networks and support vector machines methods [17]. Rativa et al. (2018) recommended different learning models including support vector regression, Gaussian process, and artificial neural networks to estimate height and weight from anthropometric measurements [5]. Jeyakumar Henry et al. (2019) used a regression equation to estimate the height and weight in Asian-Chinese adults. The arm length, knee height measurements and age are significant to estimate the height. The age, arm circumference and waist circumference are significant to estimate the weight [18]. Bhattacharjya and Kakoty (2020) focused on 72 anthropometric body dimensions, including the age and body weight in terms of gender and ethnic diversity. They used factor analysis and regression modelling to define the relations among anthropometric dimensions [19]. Son and Kim (2020) use machine learning algorithms to estimate the stature based on anthropometric data by handling missing values. They specified that support vector machine presented the highest accuracy in all ratios of missing data [20]. Wibneh et al. (2021) handled the synthesis of anthropometric diversity and workspace dimensions in ergonomic design of light armored vehicle [21]. Wang et al. (2021) introduced Generalized Regression Neural Network to predict 76 detailed body measurements from seven easily measured body features. The developed approach is more superior and easier than the current regression models [22]. Abderrahmane and Guelzim (2021) dealt with predicting the body weight based on fingerprint measurements such as fingerprint circumference, fingerprint area, fingerprint length, and fingerprint width by using more than 40 machine learning algorithms [23]. Mun et al. (2021) searched to find the association of heat and cold patterns with anthropometry/body composition. The gathered data using a self-administered questionnaire. They used a regression equation to define the correlation coefficients among variables [24]. Uçar et al. (2021) examined to determine the body fat percentage using multilayer feedforward neural networks, support vector machine regression and decision tree regression models with high accuracy rate and minimum parameter. They used age, height, weight, neck, chest, abdomen, hip, tigh, knee, ankle, biceps, forearm, wrist circumference data [25]. Jaruenpunyasak et al. (2022) handled the convolutional neural networks and traditional techniques by using the anthropometric ratios for lower-body detection [26]. Naser (2022) derived a mapping function to examine the relationship between anthropometric data and body mass index by using interpretable machine learning techniques. The author aimed to handle two goals. The first is to develop an interpretable machine learning to predict body mass index. The second is to obtain a mapping function, which shows the relationship between anthropometric data and body mass index [27]. Shi et al. (2022) handled the weight, height, body mass index, sitting height, waist-to-hip ratio, calf circumference, and 5 summary measures of limb length to predict the anthropometric measurements in 60-70-year-old women. They used on the least absolute shrinkage and selection operator regression, a machine learning approach to predict. Validating agreement was realized by using Paired t test and Bland-Altman analysis [28]. García-D'urso et al. (2022) examined the clinical and anthropometric data collected by nutritionists during dieting periods. They used a machine learning approach to predict the cholesterol levels. Different groupings of patients are identified by using a clustering analysis [29].

This study aims to develop a machine learning based model that uses weight, height, and eye height measurements as input and estimates 26 different anthropometric measurements. The Support vector regression (SVR) and Gaussian Process Regression (GPR) optimized with Bayesian algorithm based on different kernels are employed to develop estimation models. Additionally, this paper presented Bayesian optimization for hyperparameters, unlike previous studies on anthropometric measurement estimation via SVR and GPR. In this respect, it also contributes to the literature on anthropometric measurements. In addition, while weight and height variables are used as output variables in existing papers, these variables are considered as inputs in this study.

2. MATERIAL AND METHOD

2.1. Data Collection Process

This study is based on the study done with 185 male individuals between 55 and 65 years old from Ankara city of Turkey. The study subjects are selected randomly to establish cluster sampling. This study has achieved the data from patients at hospitals. Data collectors are appointed for approximately two weeks for each hospital. A total of 29 variables with three inputs and 26 outputs are collected. A Harpenden anthropometer and a digital weighing scale were used to achieve the subjects' anthropometric measurements and weight.

This study uses three anthropometric data, namely weight, stature, and eye height, as input variables to estimate various anthropometric measurements. Table 1 includes the definitions, the abbreviations (Abbr.), and the units of the inputs and outputs.

Table 1. *Definition of the Variables*

W, S, EYH are inputs used to estimate anthropometric measurements given in Table 2 that displays the descriptive statistics of inputs and outputs variables.

Table 2. *The descriptive statistics of the data set*

The histogram shows the frequency distribution of a dataset. Boxplot is used to visually show the distribution of numerical data and variability by displaying data quartiles and means. These data provide insight into a process's ability to create output. The histogram and boxplot of the input data are given in Figure 1. Histograms of the output values are given in Appendix 1.

Figure 1. The histogram and box plot of the input data.

Data preprocessing is one of the fundamental steps in the development of machine learning models. Cleaning, transforming, and modeling data is a large part of the process. Data collected from multiple sources is often found in an unorganized form, which affects the predictive performance of developed models. Therefore, raw data must go through data preprocessing before using machine learning models. Normalization of variables in multivariate analyzes is critical for accurate results, as variables measured at different scales may not contribute equally. For example, if normalization is not performed on two features in the 0-100 range and 0-1 range, the 0-100 range variable will have more weight in the model. Converting data to comparable scales can avoid this problem. This scaling can be achieved by data normalization. In the literature, two different methods, normalization, and standardization are used to bring the data to the same scale in machine learning models. Since the variables have positive values in this study, normalization was performed to bring all parameters to the same positive scale.

Bringing the variables to the same scale in the range of 0.05-0.95 provides more accurate comparisons in machine learning applications [30-31]. The data set is divided into two, 80% for training and 20% for testing. This process helps with accuracy and the learning phase efficiency. All data are normalized according to Eq. (1).

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \times 0.9 + 0.05 \quad (1)$$

where the minimum value of x is shown via x_{min} and the maximum value of x is represented $x_{\max}$ is, x_i' displays the i^{th} normalized value, and x_i depicts the i^{th} actual value. The data set was divided into two as 148 data (80%) for the training stage and 37 (20%) data for the testing stage.

This study developed estimation models using optimized SVR and GPR models using weight, stature and eye height variables as inputs to estimate various anthropometric measures. The inputs and outputs of the proposed models can be seen in Figure 2.

Figure 2. The input and output variables of the developed models

The models developed in this study are designed to estimate 26 different outputs. Figure 2 shows the inputs and outputs of the models and the developed models. Different combinations of models have been developed according to the selected inputs and outputs. For example, models developed with SVR and GPR to estimate shoulder height (SHH) by using weight (W), stature (S), eye height (EYH) input variables were optimized with the Bayesian algorithm, and the model has the best estimation accuracy selected. Therefore, the model is designated as M1. On the other hand, models developed for 26 different output variables using the same input variables are M2, M3,…, M26.

2.2. Support Vector Regression (SVR)

Support vector machine (SVM) model is used when the patterns between the input variables are indeterminate. It is a commonly preferred ML algorithm for pattern recognition and classification problems. The basis of SVM model is based on structural risk minimization. SVM differs from ML algorithms in supervised learning in that it allocates errors according to the gain of the data set, not according to the input dimensionality. Therefore, it performs well even when the dataset is extensive. The SVM-based SVR algorithm was developed due to the difficulties in adapting the method used to regression-based multi-class estimation problems [30, 32]. The purpose of SVR is to use a technique similar to the solution of regression problems in data sets with more than two variables. This way, it will be possible to calculate the regression function of data sets consisting of multidimensional feature sets. In addition, in cases where the data can be separated linearly, the data can be separated into two classes with a linear plane. However, in real-life applications, this may not always be the case. In these cases, nonlinear support vector regression is needed. SVR can solve nonlinear relationships thanks to its kernel-based structure. According to the selected kernel function, a linear or nonlinear range can be obtained when applying the SVR model. It tries to find the most appropriate regression function to represent the relationships of the data set. The mathematical expressions of SVR architecture can be seen in Eq. (2) [30, 33].

$$
\begin{gathered}
maximize \begin{cases} \frac{1}{2}\sum_{i=1}^{l}\left(w_i - w_i^*\right)\left(w_i - w_i^*\right)K\left\langle x_i, x_j \right\rangle \\ - \in \frac{1}{2}\sum_{i=1}^{l}\left(w_i + w_i^*\right) + \sum_{i=1}^{l} y_i\left(w_i - w_i^*\right) \end{cases} \\
s.t \begin{cases} \left\{ \sum_{i=1}^{l}\left(w_i - w_i^*\right) = 0 \ and \ \ w_i, w_i^* \in \left[0, C\right] \right. \\ \qquad 0 \le w_i, w_i^* \le \frac{c}{l} \\ \qquad i = 1, 2, \ldots, l \end{cases}
\end{gathered} \tag{2}
$$

In Eq. (2), ωi and ωi* are nonnegative multipliers for each observation. xi is observed data, l represents data size, C is the penalty coefficient, ϵ depicts the penalty dimension, and K (xi, xj) presents the kernel function. Therefore, it is necessary to adjust the ω to get the optimum solution. The mathematical expression of the regression equation is shown in Eq. (3) [33].

$$
f\left(x\right) = \sum_{i=1}^{l}\left(w_i - w_i^*\right)K\left(x_i - x_j\right) + b^* \tag{3}
$$

Commonly used kernel functions of SVR are shown in Table 3.

Table 3. Kernel Functions of SVR

2.3. Gaussian Process Regression (GPR)

The GPR method is a non-parametric Bayesian method. Theoretically, GPR uses an infinite number of parameters and allows the complexity level of the data to be determined based on the Bayesian approach. In this way, a relation is identified between the inputs and the outputs. Instead of the distribution of parameters for a particular function, GPR calculates the distribution for all probability functions that can describe the dataset. Therefore, the GPR model is more heuristic than other machine learning models that are sensitive to overfitting, and when estimating the mean estimate, the variance of the estimate is evaluated. This variance indicates the uncertainty in the estimates and can be precious information for specific applications. Also, GPR uses all data points and features to estimate accurately. Finally, the process of effectively optimizing the GPR model is a complex one in itself, but hyperparameter optimization improves the accuracy of the developed models [31].

Supposed that the training set T={{x_i, y_i}|i_1,2,…} is divided from the original data set, and y_i depicts a scalar target, the relationships between inputs and outputs are expressed as seen in Eq. (4) [31, 34].

$$y = x^T \beta + \varepsilon \tag{4}$$

Where ε ~N(0,σ^2) and variance of the error and β represent predicted value using the training data, in the light of Gaussian process, p(f) is zero; K is a matrix that presents a kernel function.

$$p(f) = N\left(f \mid 0, K\right) \tag{5}$$

Let K_{ij} = K(x_i,x_j), the function of y is given in Eq. (6).

$$p(y) = \int p(y \mid f)\, p(f)\, df = N\left(f \mid 0, K_y\right) \tag{6}$$

There is a latent relation f (x_i) gained for each x_i in the GPR model. Let $o_* = \left[o\left(X_*, X_1\right), ..., o\left(X_*, X_M\right)\right]^T$ and $o_{**} = k\left(x_*, x_*\right)$. Eq. (7) and Eq. (8) present the mean and variance of $P\left(y_* \mid y\right)$, respectively.

$$\mu\left(X_*\right) = o_*^T K_y^{-1} y \tag{7}$$

$$\sigma^2\left(X_*\right) = o_{**} - o_*^T K_y^{-1} o_* + \sigma_n^2 \tag{8}$$

Commonly used kernel functions of GPR are shown in Table 4.

Table 4. Kernel Functions of GPR

Kernel (Covariance) Function	**Expression**
Constant	$k = \sigma_0^2$
Linear	$k_{lin}\left(x, x'\right) = x^T x' + c$
Polynomial	$k_{poly}\left(x, x'\right) = \left(x^T x' + \sigma_0^2\right)^p$
Squared Exponential	$k_{SE}\left(r\right) = exp\left(-\frac{r^2}{2l^2}\right)$
Rational Quadratic	$k_{RQ}\left(r\right) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$
Power	$k_p\left(r\right) = -r^p$
Matern-3	$k\left(x, x'\right) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right)$

where $r = \lVert x - x' \rVert$.

2.4. The hyperparameter tuning with Bayesian optimization

Hyperparameter Optimization is the selection of suitable hyperparameters for a machine learning algorithm. The suitability of the algorithm for the dataset is related to the selection of hyperparameters. In addition, overfitting and underfitting are also directly related to this situation. Each model requires assumptions, weights, or various parameters that depend on data types under the constraints of a particular loss function. These parameters are determined by the developer in classical machine learning problems. However, the selection of the best hyperparameters is a primary complex problem. Scanning the entire solution space and selecting the most appropriate hyperparameter is possible with optimization algorithms [35]. For this reason, hybrid models have been used recently to increase the accuracy of machine learning algorithms. This study used GPR and SVR algorithms to estimate 26 different anthropometric measurements using the same input variables. The hyperparameters such as kernel function, box constraint, kernel scale, epsilon, the length scale parameter (σL), and the signal standard deviation (σF) are determined by Bayesian optimization. Firstly, the most proper kernel function was determined by Bayesian optimization, and then the parameters of the kernel function were selected. Bayesian optimization creates a probability model of the objective function and uses selecting the hyperparameter to appraise the actual objective function. Two methods frequently used for hyperparameter optimization in the literature are Grid search and Random search. In Bayesian Optimization, the performance of past hyperparameters affects the future decision. In contrast, new hyperparameters in Random Search and Grid Search algorithms are not affected by historical performance. Therefore, Bayesian Optimization is a much more robust method [36].

2.5. Performance evaluation criteria

Statistical performance metrics were used to evaluate the estimation performance of developed ML algorithms models to estimate anthropometric measurements. Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) are the error measurement statistics frequently used in previous studies [37]. These four criteria were calculated to compare the ability of the developed models in this study. Performance measures can be calculated using equations (9) to (11).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_i^{obs} - y_i^{est}\right| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i^{obs} - y_i^{est})^2} \tag{10}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_i^{obs} - \overline{y_i^{obs}})\,(y_i^{est} - \overline{y_i^{pre}})}{\sqrt{\sum_{i=1}^{n}(y_i^{obs} - \overline{y_i^{obs}})^2(y_i^{est} - \overline{y_i^{pre}})^2}}\right)^2 \tag{11}$$

where the number of observations represents n, y_i^{obs} is the observed value of the anthropometric measurements and y_i^{est} is the estimated value of the anthropometric measurements at the time i.

3. RESULTS AND DISCUSSION

3.1. Results of optimized SVR and GPR

The SVR model is designed to identify the relationship between three inputs (weight, height, eye stature) and 26 outputs explained in Table 5. SVR model includes hyperparameters, namely Kernel function, box constraint, kernel scale, and epsilon, which affect the model's predictive performance. This study developed models by optimizing these hyperparameters using Bayesian optimization to estimate 26 different outputs using three input variables. The coding for the hyperparameter optimization process, which includes kernel selection and parameter optimization, was employed using Matlab 2020a software.

While developing the SVR model, the functional connection between the inputs and outputs is determined during the training phase, and the hyperparameter values with the smallest error value are selected. Finally, the selected estimation model is evaluated in the testing phase. In addition, the selection of the kernel function and the adjusting of the parameters of the selected kernel function are optimized using Bayesian optimization. As a result, each developed model has specific hyperparameter values. As in the SVR model development process, all the necessary steps for kernel determination and hyperparameters adjustment while designing the GPR model were developed with Matlab 2020a software. The GPR has a kernel that determines the distribution's covariance for the output variable, and the appropriate probability function is determined using the training data.

The estimation accuracy of GPR models is directly dependent on the choice of kernel function and hyperparameters. The hyperparameters optimized in the GPR model are the length scale parameter (σL) and the signal standard deviation (σF). With Bayesian optimization, firstly, the kernel function is selected, and then the parameter values of the selected function are optimized. Finally, the nonlinear exponential model is solved by the Quasi-Newton method. In addition, 5-fold cross-validation was used in the training phase to prevent overfitting in each developed model. On the other hand, a random split entails dividing the data into a training set and a validation set, with a fixed proportion of the data (e.g., 80/20) assigned to each. While a random split is easier and faster to execute, it can result in unreliable estimates of model performance if the split is not representational.

Cross-validation is preferred by some researchers over random splits because it gives a more reliable estimate of a model's success on new, unseen data. Cross-validation divides data into numerous folds and uses each fold as a validation set while the remaining folds are used for training. This procedure is repeated several times, with the results averaged to obtain a more reliable estimate of the model's performance. Cross-validation and random division can both be applied to the same model. For example, cross-validation can be used to tune the model's hyperparameters before using a random split to receive a final estimate of the model's performance. Alternatively, cross-validation can be used to estimate the model's performance, and then a random split can be used to validate the model's performance on a totally new dataset. The decision to use cross-validation, a random split, or a combination of both is determined by the particular problem, the size of the dataset, and the available resources. The cases where both methods are used together are explained below.

Hyperparameter tuning: When training a machine learning model, tuning the values of hyperparameters is often necessary to optimize the model's performance. One common approach is cross-validation to evaluate the model's performance for different hyperparameter settings. After determining the optimal hyperparameters using cross-validation, a random split can be used to obtain a final estimate of the model's performance on new data [38].

Final model evaluation: Once the model's hyperparameters have been tuned, obtaining a final estimate of the model's performance on new data is important. In this case, cross-validation can be used to obtain an initial estimate of the model's performance. A random split can validate the model's performance on a new dataset. This can help to ensure that the model is not overfitting to the training data and can generalize well to new, unseen data [39].

Limited data availability: In some cases, the dataset may be small, and it may not be possible to set aside a large portion of the data for either cross-validation or a random split. In this case, it may be beneficial to combine both methods. For example, one could perform cross-validation using a smaller subset of the data and then use a random split on the remaining data to obtain a final estimate of the model's performance [40].

Whether to use cross-validation, a random split, or a combination of both depends on the specific problem, the size of the dataset, and the resources available. The key is to ensure that the model's performance is evaluated robustly and reliably, considering the data's limitations and constraints and the available computational resources. Due to the hyperparameter optimization applied in this study and the inadequacy of the data, the two methods were applied together.

3.2. Accuracy evaluations of the developed models

In this paper, several anthropometric measures are target values, and three indicators are inputs for all developed models. All models were obtained using the training datasets of different combinations of the regression relations that best describe the relationship between the inputs and the output. The models that describe these relationships developed SVR and GPR algorithms using Bayesian optimization, and test data sets were used to measure the estimation accuracy. Finally, estimation results were compared based on three metrics, namely MAE, RMSE, and MSE. The accuracy of all models can be seen in Table 5. Also, the developed models were ranked via RMSE values, and the best model was selected for each combination.

Table 5. The accuracy of the developed models for testing phases

All models successfully estimate anthropometric measurements using the specified indicators. As seen in Table 5, BO-GPR model has better accuracy than BO-SVR for all combinations except M18. BO-SVR and BO-GPR models are developed with quadratic kernel function Matern 5/2 kernel function, respectively. The RMSE values of the best models selected for each combination varied between 0.255 and 0.319 during the testing phase. These results show that the developed estimation models can be used when anthropometric measurements are not always possible. Especially the estimations made with BO-GPR have a shallow error rate.Table 5 shows that the developed BO-GPR model has more strong estimation ability than BO-SVR models except for M18. Figure 3 compares developed models based on RMSE. Again, only BO-SVR for M18 has a lower RMSE value than BO-GPR for all combinations. Figure 4 compares the estimated values with the best accuracy for all combinations to the observed values.

Figure 3. The comparison of developed models via RMSE

Figure 4. Observed and Estimated values for developed models

4. Conclusion

ML-based prediction models can uncover patterns and relationships in large datasets that may not be immediately apparent to humans. In some cases, these models can reveal previously unknown relationships or correlations that can lead to new insights and discoveries. However, it's important to note that ML-based models do not replace domain knowledge and human expertise. The insights obtained from these models must be carefully analyzed and interpreted by domain experts to ensure that they are accurate and meaningful. ML models may also uncover spurious correlations or false positives that need to be carefully evaluated to avoid drawing incorrect conclusions. ML-based models complement and enhance domain knowledge by providing a more comprehensive and data-driven problem understanding. They can help identify key factors and predictors of a particular outcome or event and can be used to develop more accurate and effective interventions or strategies. However, it's important to approach these models cautiously and carefully evaluate their results in the context of existing domain knowledge and expertise. The anthropometric measurements and weights of 185 men aged 55 to 65 from Ankara, Turkey, were collected for this research. The respondents provide 29 variables, each with three inputs and twenty-six outputs. In this study, machine learning-based models were developed to estimate anthropometric measurements using weight, height, and eye height. Thus, using these models, other anthropometric measurements of employees whose weight, height, and eye height are measured can be obtained.

In this study, machine learning regression models present better results than traditional statistical regressions to predict the anthropometric measurements from weight, stature, and eye height. This study handles SVR optimized with Bayesian based on quadratic kernel, GPR optimized with Bayesian based on matern5/2 kernel as machine learning regression models. This paper applies Bayesian models in machine learning methods to optimize the parameters as a difference from the literature. These two methods are implemented to the anthropometric measurements the first time in this paper. The estimation results are compared based on three metrics, which are MSE, RMSE, MAE. GPR optimized with Bayesian model has better accuracy than SVR optimized with Bayesian for all combinations except interpupillary distance according to the obtained results. Future papers can be focused on applying the developed machine learning methods in different areas such as product design, energy estimation, and the heuristic optimization methods can be used to optimize hyperparameters.

References

[1] World Health Organization and others. "Physical status: The use and interpretation of anthropometry", *WHO Tech. Rep. Ser.*, Geneva, vol. 854, pp. 2009 (1995).

[2] Bloomfield, R., Steel, E., MacLennan, G. et al. "Accuracy of weight and height estimation in an intensive care unit: Implications for clinical practice and research", *Critical care medicine*, *34*(8), pp. 2153-2157 (2006).

[3] Anglemyer, B. L., Hernandez, C., Brice, J. H. et al. "The accuracy of visual estimation of body weight in the ED", *The American journal of emergency medicine*, *22*(7), pp. 526-529 (2004).

[4] Friedman, J. H. "Data Mining and Statistics: What's the connection?", *Computing science and statistics*, *29*(1), pp. 3-9 (1998).

[5] Rativa, D., Fernandes, B. J. and Roque, A. "Height and weight estimation from anthropometric measurements using machine learning regressions", *IEEE journal of translational engineering in health and medicine*, 6, pp. 1-9 (2018).

[6] Lisboa, P. J. and Taktak, A. F. "The use of artificial neural networks in decision support in cancer: a systematic review", *Neural networks*, *19*(4), pp. 408-415 (2006).

[7] Bishop, C. M. "Pattern Recognition and Machine Learning", New York, NY, USA: Springer-Verlag. pp. 325-355 (2016).

[8] Schölkopf, B., Burges, C. J. C. and Smola A. J., *Advances in Kernel Methods_Support Vector Learning*. Cambridge, MA, USA: MIT Press, pp. 1-17 (1999).

[9] Shawe-Taylor, J. and Sun, S., "Kernel methods and support vector machines" in *Academic Press Library in Signal Processing*, vol. 1. New York, NY, USA: Elsevier, pp. 857_881 (2014).

[10] Chumlea, W. C. and Guo, S. "Equations for predicting stature in white and black elderly individuals", *Journal of Gerontology*, 47(6), pp. 197-203 (1992).

[11] Michels, K. B., Greenland, S. and Rosner, B. A. "Does body mass index adequately capture the relation of body composition and body size to health outcomes?", *American Journal of Epidemiology*, 147(2), 167-172 (1998).

[12] Kaya, M. D., Hasiloglu, A. S., Bayramoglu, M. et al. "A new approach to estimate anthropometric measurements by adaptive neuro-fuzzy inference system", *International Journal of Industrial Ergonomics*, 32(2), pp. 105-114 (2003).

[13] Gauld, L. M., Kappers, J., Carlin, J. B. et al. "Height prediction from ulna length", *Developmental medicine and child neurology*, 46(7), pp. 475-480 (2004).

[14] Hu, H., Li, Z., Yan, J. et al. "Anthropometric measurement of the Chinese elderly living in the Beijing area", *International Journal of Industrial Ergonomics*, 37(4), 303-311 (2007).

[15] Kuiti, B. and Bose, K. "Predictive equations for height estimation using knee height of older Bengalees of Purba Medinipur, West Bengal, India", *AnthropologicAl review*, 79(1), pp. 47-57 (2016).

[16] Lee, B. J., Lee, J. C., Nam, J. et al. "Prediction of cold and heat patterns using anthropometric measures based on machine learning", *Chinese Journal of Integrative Medicine*, 24(1), pp. 16-23 (2018).

[17] Ferenci, T. and Kovacs, L. "Predicting body fat percentage from anthropometric and laboratory measurements using artificial neural networks", *Applied Soft Computing*, 67, pp. 834-839 (2018).

[18] Jeyakumar Henry, C., Ponnalagu, S., and Bi, X. "Equations to predict height and weight in Asian-Chinese adults", *Malaysian Journal of Nutrition*, 25(3) pp. 393-403 (2019).

[19] Bhattacharjya, B. R. and Kakoty, S. K. "A survey of the anthropometric data relating to five ethnic groups in Assam considering gender and ethnic diversity: Application of the data in designing an improvised pedal-operated Chaak", *International Journal of Industrial Ergonomics*, 76, pp.1-21 (2020).

[20] Son, Y. and Kim, W. "Missing Value Imputation in Stature Estimation by Learning Algorithms Using Anthropometric Data: A Comparative Study", *Applied Sciences*, 10(14), 5020 pp. 1-13 (2020).

[21] Wibneh, A., Singh, A. K. and Karmakar, S. "Understanding the synthesis of anthropometric diversity and workspace dimensions in ergonomic design of light armored vehicle", *Human Factors and Ergonomics in Manufacturing & Service Industries*, *31*(5), pp. 447-468 (2021).

[22] Wang, L., Lee, T. J., Bavendiek, J. et al. "A data-driven approach towards the full anthropometric measurements prediction via Generalized Regression Neural Networks", *Applied Soft Computing*, 109, 107551 (2021)

[23] Abderrahmane, M. A. and Guelzim, I. "Comprehensive Study on Body Weight Estimation Based on Fingerprint Measurements using Machine Learning", I*n 2021 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 173-177). IEEE (2021).

[24] Mun, S., Park, K. and Lee, S. "Study on the Anthropometric and Body Composition Indices for Prediction of Cold and Heat Patter", J*ournal of Korean Medicine*, 42(4), pp. 185-196 (2021).

[25] Uçar, M. K., Ucar, Z., Köksal, F. et al. "Estimation of body fat percentage using hybrid machine learning algorithms", *Measurement*, 167, 108173 (2021).

[26] Jaruenpunyasak, J., García Seco de Herrera, A. and Duangsoithong, R."Anthropometric ratios for lower-body detection based on deep learning and traditional methods", *Applied Sciences*, *12*(5), 2678 (2022).

[27] Naser, M. Z. "Deriving mapping functions to tie anthropometric measurements to body mass index via interpretable machine learning", *Machine Learning with Applications*, 8, 100259, pp. 1-7 (2022).

[28] Shi, J., He, Q., Pan, Y. et al. "Estimation of Appendicular Skeletal Muscle Mass for Women Aged 60-70 Years Using a Machine Learning Approach", *Journal of the American Medical Directors Association*, 23(12), 1985.e1-1985.e7 (2022)..

[29] García-D'urso, N., Climent-Pérez, P., Sánchez-Sansegundo, M. et al. "A non-invasive approach for total cholesterol level prediction using machine learning", *IEEE Access*, *10*, pp. 58566-58577 (2022).

[30] Guleryuz, D. "Predicting Health Spending in Turkey Using the the GPR, SVR, and DT Models", *Acta Infologica*, *5*, pp. 155–166 (2021).

[31] Ozden, E. and Guleryuz, D. "Optimized machine learning algorithms for investigating the relationship between economic development and human capital", *Computational Economics*, *60*(1), pp. 347-373 (2022).

[32] Wang, J., Lei, C. and Guo, M. "Daily natural gas price forecasting by a weighted hybrid data-driven model", *Journal of Petroleum Science and Engineering*, *192*(December 2019), 107240 (2020).

[33] Quan, Q., Hao, Z., Xifeng, H. et al. "Research on water temperature prediction based on improved support vector regression", *Neural Computing and Applications*, 34, pp. 8501–8510 (2022).

[34] Zhang, C., Wei, H., Zhao, X. et al. "A Gaussian process regression based hybrid approach for short-term wind speed prediction", *Energy Conversion and Management*, *126*, pp. 1084–1092 (2016).

[35] Liashchynskyi, P. and Liashchynskyi, P. "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS", *arXiv preprint arXiv:1912.06059*, pp. 1–11 (2019).

[36] Bergstra, J., Bardenet, R., Bengio, Y. et al. "Algorithms for hyper-parameter optimization", *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*, *NIPS 2011*, pp. 1–9 (2011).

[37] Chaibakhsh, A. "Modelling and long-term simulation of a heat recovery steam generator", *Mathematical and Computer Modelling of Dynamical Systems*, 19(2), pp. 91–114 (2013).

[38] Bischl, B., Binder, M., Lang, M. et al. "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1484 (2021).

[39] Sun, H., Burton, H. V. and Huang, H. “Machine learning applications for building structural design and performance assessment: State-of-the-art review”, *Journal of Building Engineering*, 33, 101816 (2021).

[40] Vabalas A, Gowen E, Poliakoff E.et al. “Machine learning algorithm validation with a limited sample size”, *PLoS ONE* 14(11): e0224365. (2019).

Table 1. Definition of the Variables

Table 2. The descriptive statistics of the data set

Table 3. Kernel Functions of SVR

Table 4. Kernel Functions of GPR

Table 5. The accuracy of the developed models for testing phases

Figure 1. The histogram and box plot of the input data.

Figure 2. The input and output variables of the developed models

Figure 3. The comparison of developed models via RMSE

Figure 4. Observed and Estimated values for developed models

Table 1. Definition of the Variables

Variable	Abbr.	Unit	Variable	Abbr.	Unit
Weight	W	Kilogram	Ankle height (M13)	ANH	millimeter
Stature	S	Millimeter	Functional thumb-tip reach (M14)	FTR	millimeter
Eye height	EYH	Millimeter	Popliteal height (M15)	POH	millimeter
Shoulder height (M1)	SHH	Millimeter	Maximum head breadth (M16)	MHB	millimeter
Middle fingertip height (M2)	MFH	Millimeter	Maximum head length (M17)	MHL	millimeter
Waist height (M3)	WAH	Millimeter	Interpupillary distance (M18)	IND	millimeter
Elbow height (M4)	ELH	Millimeter	Total head height (M19)	THH	millimeter
Functional hand height (M5)	FHH	Millimeter	Maximum handbreadth (M20)	MHN	millimeter
Tibial height (M6)	TIH	Millimeter	Hand length (M21)	HAL	millimeter
Crotch height (M7)	CRH	Millimeter	Finger length (M22)	FIL	millimeter
Shoulder breadth (M8)	SHB	Millimeter	Foot length (M23)	FOL	millimeter
Hip breadth (M9)	HIB	Millimeter	Foot breadth (M24)	FOB	millimeter
Waist depth (M10)	WAD	Millimeter	Forearm-fingertip length (M25)	FFL	millimeter
Waist breadth (M11)	WAB	Millimeter	Buttock-knee length (M26)	BKL	millimeter
Thigh circumference (M12)	THC	Millimeter			

Table 2. The descriptive statistics of the data set

	W	S	EYH	SHH	MFH	WAH	ELH	FHH	TIH	CRH	SHB	HIB	WAD	WAB	THC
Minimum	57	1385	1293	1152	525	835	856	608	392	603	276	288	216	259	423
Maximum	79	1924	1796	1600	729	1160	1189	844	544	837	384	400	300	359	587
Mean	69	1670	1559	1389	633	1007	1032	733	472	727	333	347	260	312	510
Std	7	161	150	134	61	97	99	70	45	70	32	33	25	30	49
Kurtosis	-1.22	-1.22	-1.23	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22	-1.22
Skewness	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12
	ANH	**FTR**	**POH**	**MHB**	**MHL**	**IND**	**THH**	**MHB**	**HAL**	**FIL**	**FOL**	**FOB**	**FFL**	**BKL**	
Minimum	55	638	337	131	157	51	190	85	150	58	203	76	381	460	
Maximum	77	886	469	181	217	71	264	119	208	80	281	106	529	639	
Mean	67	769	407	157	189	62	229	103	181	70	244	92	459	555	
Std	6	74	39	15	18	6	22	10	17	7	23	9	44	53	
Kurtosis	-1.23	-1.22	-1.22	-1.22	-1.22	-1.20	-1.21	-1.22	-1.23	-1.21	-1.22	-1.22	-1.22	-1.22	
Skewness	-0.10	-0.12	-0.12	-0.12	-0.12	-0.12	-0.11	-0.12	-0.12	-0.12	-0.12	-0.11	-0.12	-0.12	

Table 3. Kernel Functions of SVR

Kernel Function	Expression	Parameters
Linear	$K(x_i, x_j) = (x_i, x_j)$	
Polynomial	$K(x_i, x_j) = (\langle x_i, x_j \rangle + 1)^d$	D
Gaussian	$K(x_i, x_j) = e^{(-\frac{\|x_i - x\|^2}{2\gamma^2})}$	γ
Sigmoid	$K(x_i, x_j) = tanh(\gamma \langle x_i, x_j \rangle + 1)^d$	γ, d

Table 4. Kernel Functions of GPR

Kernel (Covariance) Function	**Expression**
Constant	$k = \sigma_0^2$
Linear	$k_{lin}(x, x') = x^T x' + c$
Polynomial	$k_{poly}(x, x') = (x^T x' + \sigma_0^2)^p$
Squared Exponential	$k_{SE}(r) = exp\left(-\frac{r^2}{2l^2}\right)$
Rational Quadratic	$k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$
Power	$k_p(r) = -r^p$
Matern-3	$k\,(x, x') = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right)$

where $r = \|x - x'\|$.

Table 5. The accuracy of the developed models for testing phases

Input-Output Combinations	**BO-SVR**			**BO-GPR**		
	MAE	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
M1	8.738	100.774	10.039	**0.285**	**0.102**	**0.319**
M2	6.323	43.491	6.595	**0.252**	**0.087**	**0.294**
M3	6.011	48.141	6.938	**0.267**	**0.092**	**0.303**
M4	6.232	52.154	7.222	**0.234**	**0.077**	**0.277**
M5	4.564	27.779	5.271	**0.248**	**0.082**	**0.286**
M6	2.764	10.195	3.193	**0.238**	**0.080**	**0.283**
M7	4.369	25.323	5.032	**0.213**	**0.065**	**0.255**
M8	2.053	5.595	2.365	**0.239**	**0.076**	**0.276**
M9	2.197	6.280	2.506	**0.273**	**0.097**	**0.311**
M10	0.378	0.179	0.424	**0.267**	**0.100**	**0.317**
M11	1.927	4.854	2.203	**0.239**	**0.079**	**0.282**
M12	3.100	13.117	3.622	**0.252**	**0.084**	**0.289**
M13	0.355	0.180	0.425	**0.238**	**0.075**	**0.273**
M14	4.860	30.846	5.554	**0.223**	**0.067**	**0.260**
M15	2.448	7.981	2.825	**0.239**	**0.083**	**0.288**
M16	1.002	1.234	1.111	**0.255**	**0.095**	**0.308**
M17	1.248	1.977	1.406	**0.255**	**0.086**	**0.292**
M18	**0.252**	**0.088**	**0.296**	0.285	0.104	0.323
M19	1.659	3.445	1.856	**0.263**	**0.098**	**0.314**
M20	0.316	0.154	0.392	**0.253**	**0.087**	**0.295**
M21	1.121	1.737	1.318	**0.216**	**0.068**	**0.261**
M22	0.386	0.209	0.458	**0.226**	**0.077**	**0.278**
M23	1.648	3.388	1.841	**0.269**	**0.091**	**0.301**
M24	0.507	0.360	0.600	**0.265**	**0.096**	**0.310**
M25	3.290	13.487	3.673	**0.269**	**0.095**	**0.308**

M26	3.508	16.536	4.066	**0.225**	**0.072**	**0.267**

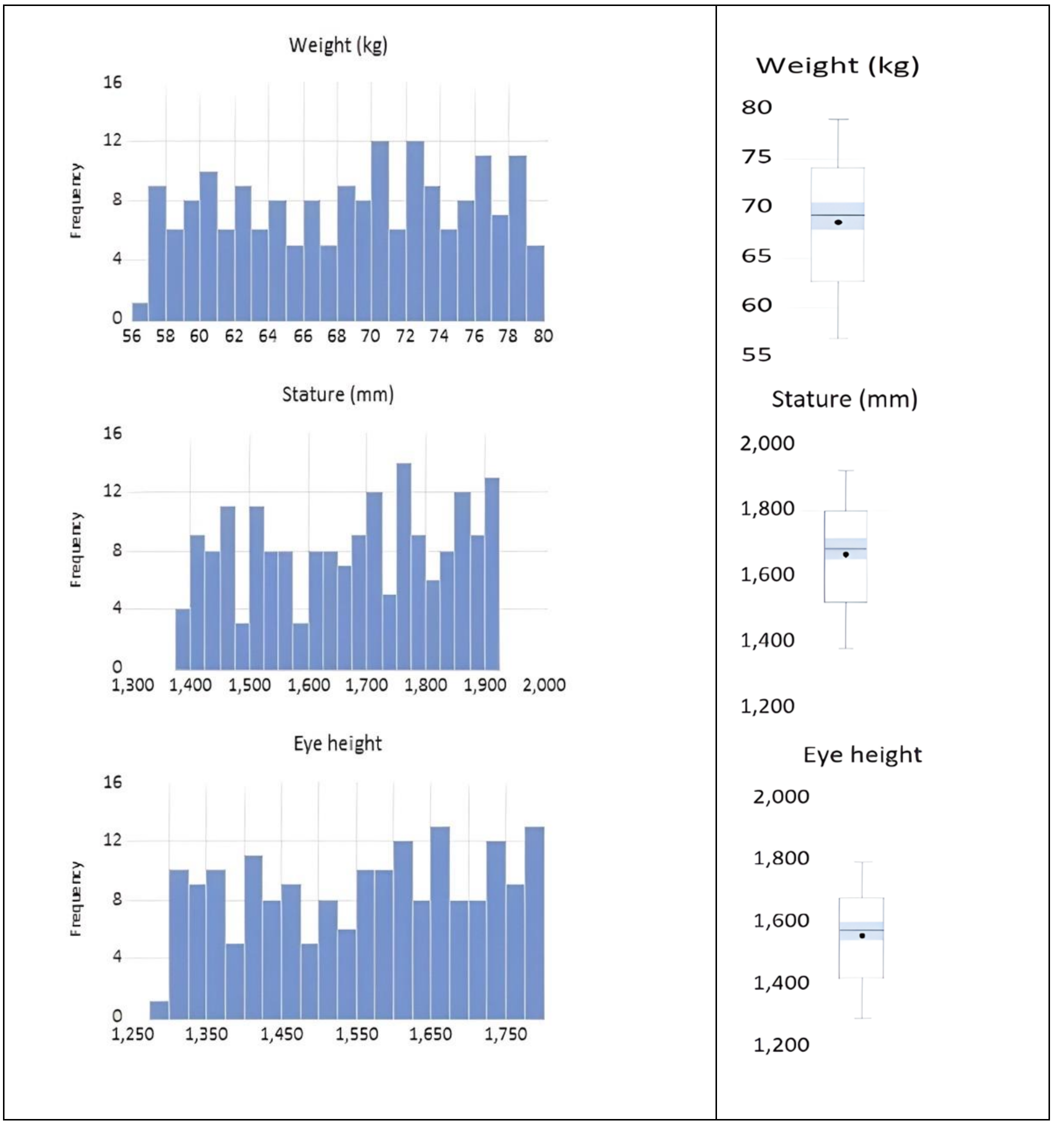
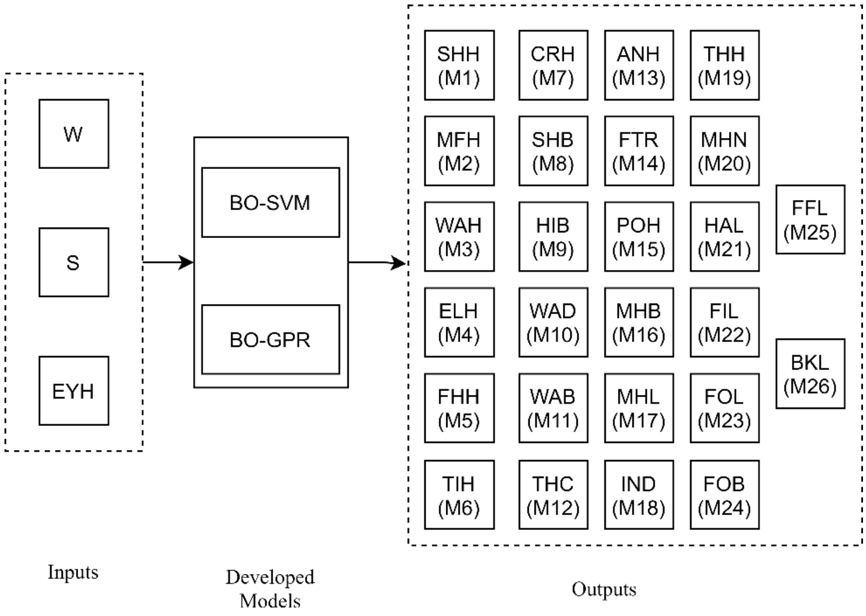


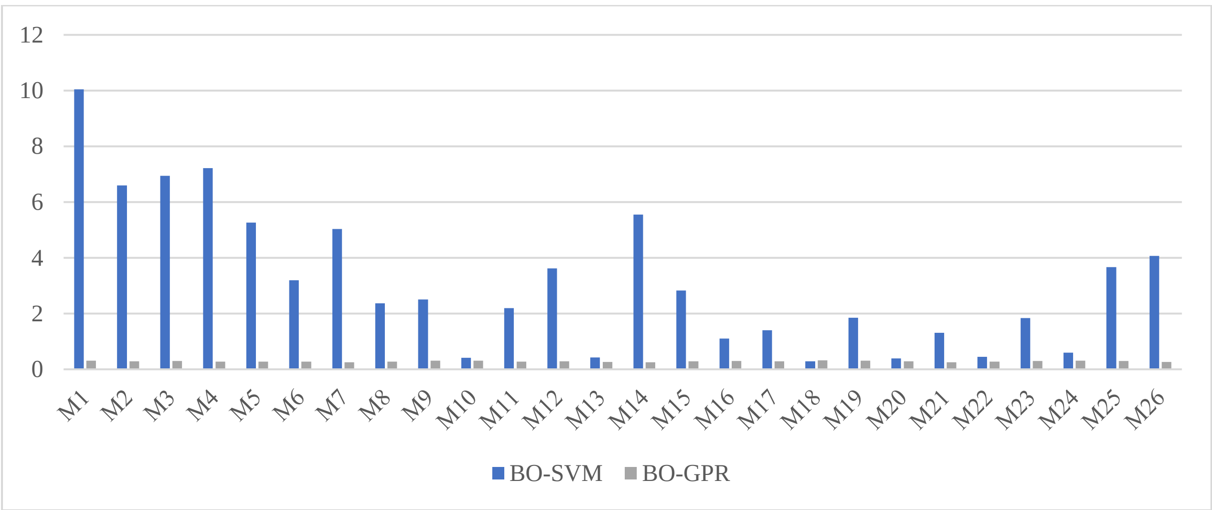
Figure 1. The histogram and box plot of the input data.



516

517 **Figure 2.** The input and output variables of the developed models

518



519

Figure 3. The comparison of developed models via RMSE

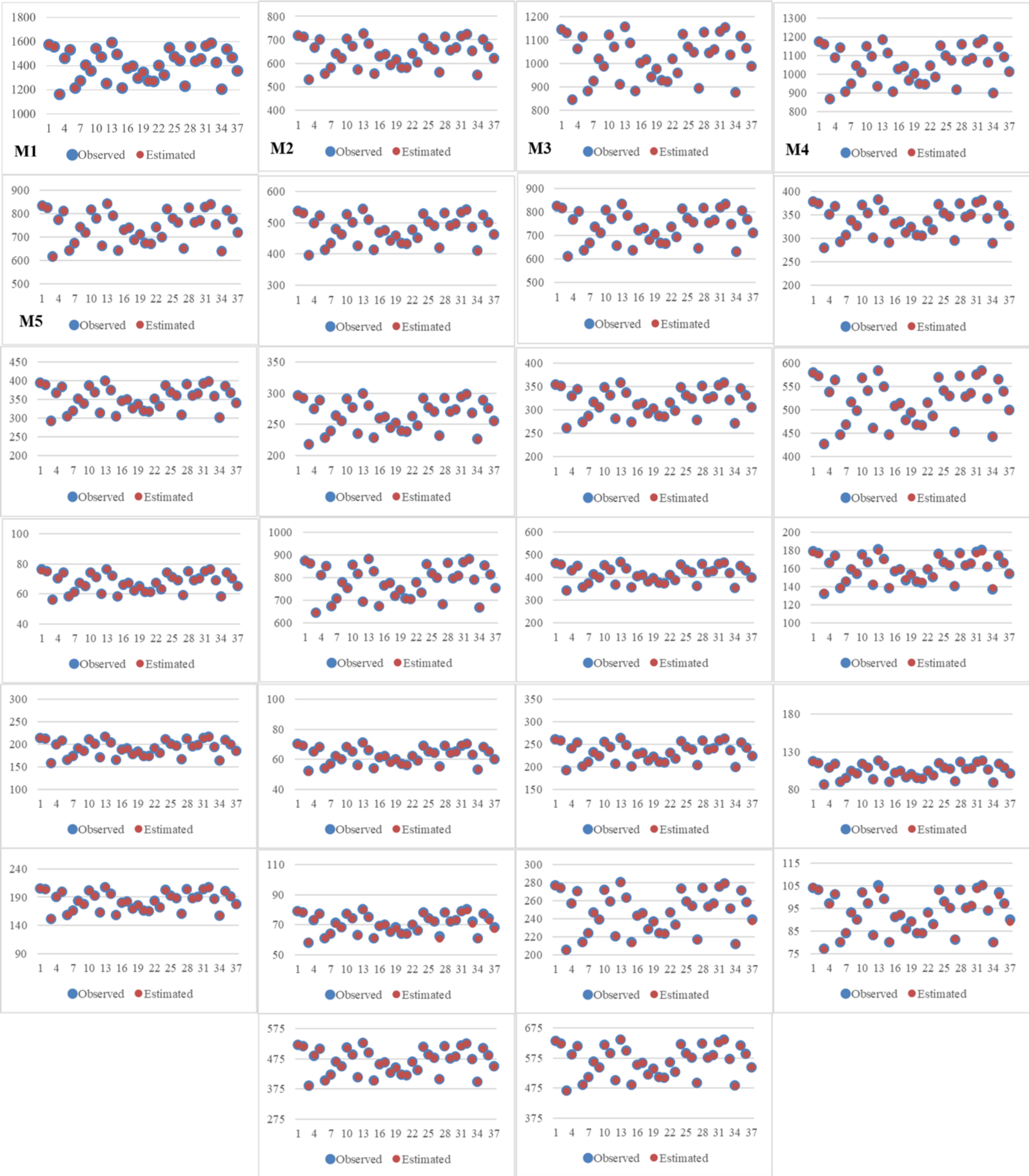
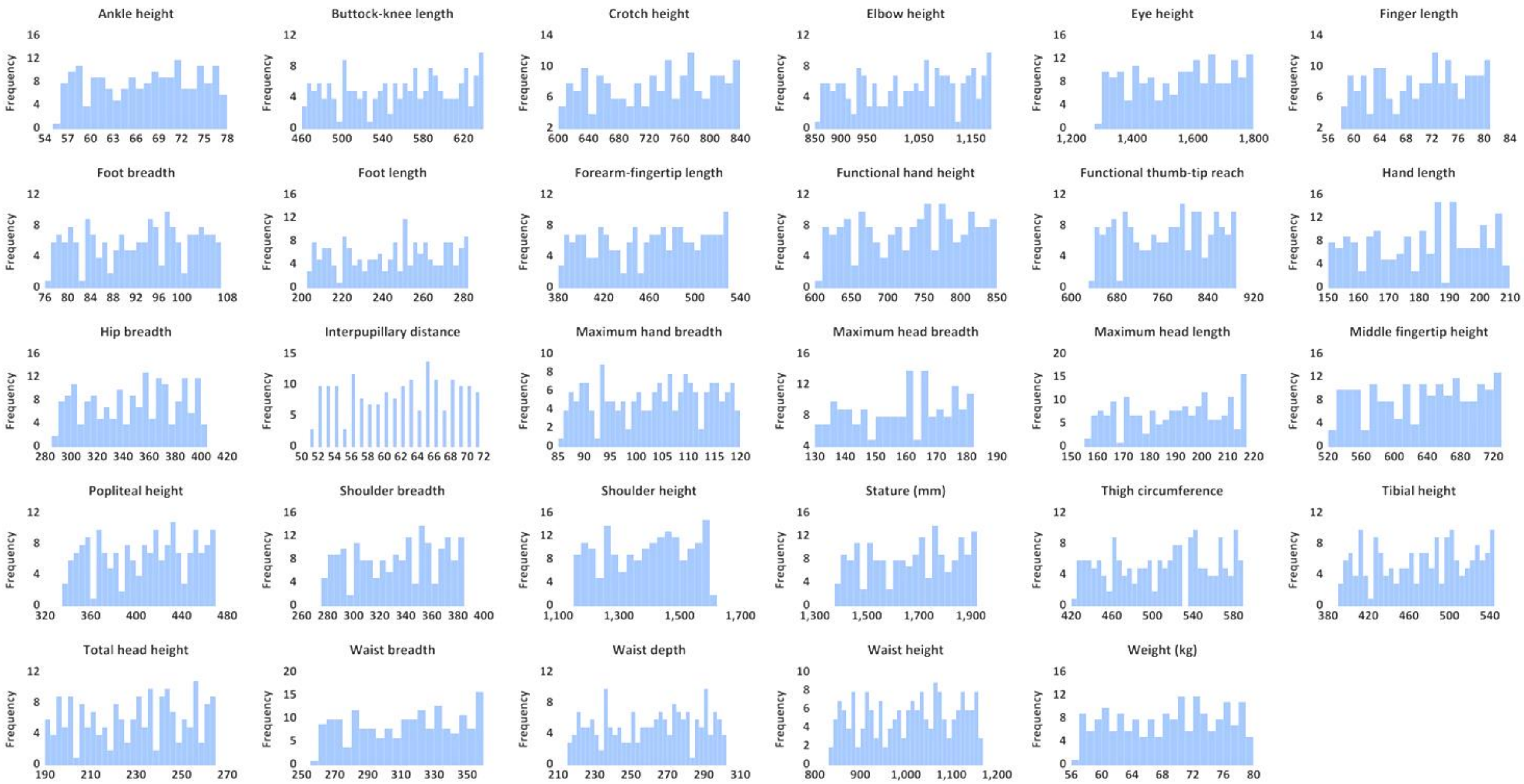


Figure 4. Observed and Estimated values for developed models

APPENDIX 1. Histograms of output variables.



Biographical notes:

Didem Güleryüz is an Assoc. Prof. Dr. at the Department of Management Information Systems, Bayburt University, Turkey. He received her PhD in Industrial Engineering at İstanbul University, Turkey. His current research interests include machine learning and artificial intelligence.

Ömer Faruk Efe graduated from Selçuk University, Department of Industrial Engineering in 2008. He received an M.Sc. degree in Industrial Engineering from Selçuk University. He received a Ph.D. degree in Industrial Engineering from Sakarya University. His research interests are Multi-criteria decision making, fuzzy logic, lean production, ergonomics, occupational health, and safety. He worked Gümüşhane University and Afyon Kocatepe University. He has been working as an Associate Professor at Bursa Technical University. He has published various research papers in international/national journals. He has studied on some book chapters.

Burak Efe is an Associate Professor at the Department of Industrial Engineering, Necmettin Erbakan University, Turkey. He received his MSc and PhD in Industrial Engineering at Gazi University, Turkey. He also worked in Gazi University. He have lectured many courses such as human factors, work analysis and design, production planning. He has published various research papers in international/national journals. He has studied on some book chapters. His current research interests include assembly line balancing, ergonomics, fuzzy logic, multi-criteria decision making, risk assessment.