STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LX**, Number 2, 2015

APPLYING SUPPORT VECTOR REGRESSION METHODS FOR HEIGHT ESTIMATION IN ARCHAEOLOGY

VLAD-SEBASTIAN IONESCU

Abstract. In this paper we apply Support Vector Machines to the problem of predicting the height of human skeletons given bone measurements. There exist archaeological methods for estimating height, but our purpose is to investigate the potential of Support Vector Regression for this task. Since skeletal stature clearly depends on individual bone lengths, building SVM models for this task has the potential of giving an accurate machine learning automation for this task, which can be useful for archaeologists. We investigate multiple kernels and performance evaluation methodologies and compare our results to existing literature results on the topic. Our experiments show that SVM regression models are very good for the problem at hand, outperforming existing approaches.

1. Introduction

A very important problem in archaeology and forensic science is the problem of height estimation. Existing approaches that deal with this problem involve simple regression formulas based on statistical methods. Our goal is to investigate the potential of more complex methods, such as Support Vector Machines (SVMs). We believe that the good performance of Support Vector Machines on other problems can make them ideal for solving the problem of height estimation as well, due to their good performance at inferring complex relationships between data: in our case between the bone measurements and the associated height of the skeletons.

This is a problem that is difficult even for humans, with no clear relationship between the features and their labels, which suggests the use of machine learning models.

Received by the editors: June 3, 2015.

2010 *Mathematics Subject Classification.* 68T05, 68T01.

1998 *CR Categories and Descriptors.* I.2.6 [**Artificial Intelligence**]: Learning – *Concept learning*; I.2.7 [**Artificial Intelligence**]: Learning – *Induction*.

Key words and phrases. height estimation, regression, support vector machines, archaeology, forensic science.

The problem is important for researchers in archaeology and related fields because it allows them to discover important facts about a certain population, relating to issues such as their health, gender differences and body sizes at different times in history.

We propose using SVM regression models because of their known good performance in general on regression tasks, their efficient implementations in various libraries and their ability to easily adapt to multiple problem settings through the use of kernels. To the best of our knowledge, our approach is novel, since SVM models have not been used in the literature for height estimation until now. Our obtained experimental results are significantly better than the existing ones in the literature and prove the ability of machine learning models and SVMs in particular to solve this problem.

Our experiments are performed on open source skeletal data which was previously used for this task. Because of this, we have a relevant baseline to compare our results against.

The paper has the following structure. Section 2 presents the problem of height estimation and its importance for archaeologists, as well as existing approaches for solving it. Section 3 provides an overview of Support Vector Regression. Section 4 presents our experimental setup and methodology. Section 5 presents our data sets and the results obtained on each one. Section 6 presents a comparison of our results to related work and Section 7 contains our conclusions and future research directions.

2. The height estimation problem

According to [17, 3], height estimation is a central part of anthropological analysis and is generally used in order to determine social stucture in extinct populations. More complex theories can also be inferred from stature, relating to health, body size trends and adaptability to environmental changes.

The first anatomical method for height estimation was introduced by Thomas Dwight in 1894, a method that caused significant errors. Karl Pearson then introduced regression formulas, but there were studies that identified certain shortcomings regarding their applicability to different populations [10, 8, 9].

A milestone approach is introduced in [2], along with the open source data sets we use in this article. The approach consists of multiple formulas based on measurements of important bones in the human body.

A variety of approaches exist for this problem, which proves its importance. The existing methods are either anatomical or mathematical in nature. A study comparing the two classes of methods can be found in [18], which concludes that anatomical methods are superior only if the skeletal remains

are sufficiently complete and that otherwise the mathematical methods are to be preferred.

A comprehensive literature review on this topic can be consulted in [15]. We have previously introduced two novel machine learning models based on artificial neural networks and genetic algorithms for the problem approached in this paper [7].

3. Support Vector Regression

Support Vector Machines were introduced for binary classification by Vapnik as far back as 1963 and further developed by Cortes and Vapnik [16]. The idea behind SVMs is to find a maximum margin separating hyperplane. They are very robust to different problems due to the kernel trick, which allows them to accurately do non-linear classification as well, and due to the fact that, unlikely neural networks, they do not have to concern themselves with local optimums.

Support Vector Regression is an extension of SVMs to regression problems. A widely used method is called ε-Support Vector Regression, in which we want to find a function that approximates each training example with at most ε error, if possible, or allow for some degree of error (specified by a hyperparameter $C > 0$) if not. This is similar to the width of the margin in Support Vector Classification.

The mathematical formulation in which only an error of ε is allowed is given in Formula (1) [16, 1].

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 \\ &\text{subject to} \quad \begin{cases} y_i - (w \cdot x_i + b) \leq \varepsilon \\ (w \cdot x_i + b) - y_i \leq \varepsilon \end{cases} \end{aligned} \tag{1}$$

The more robust formulation that allows for some mistakes and is used in practice is given in Formula (2) [16, 1].

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i^- + \xi_i^+) \\ &\text{subject to} \quad \begin{cases} y_i - (w \cdot x_i + b) \leq \varepsilon + \xi_i^- \\ (w \cdot x_i + b) - y_i \leq \varepsilon + \xi_i^+ \\ \xi_i^-, \xi_i^+ \geq 0 \end{cases} \end{aligned} \tag{2}$$

In both equations, standard notation is used: w is a weight vector that we use to multiply the input with and make predictions, m is the number

of training instances, x_i is a training sample, y_i is its target and b is a bias term. The resulting optimization problems are then solved using numerical algorithms. Various kernel functions can also be used in order to obtain non-linear classification or regression [16].

Intuitively, Support Vector classification and regression are similar in that classification attempts to find a separating hyperplane that goes as much through the "middle" of the space that exists between the data classes and regression attempts to find a hyperplane that goes as much through the "middle" of the data as possible. The hyperparameter C is what controls how much "slack" we give the algorithm, or how much we tolerate mistakes.

For reference, Formula (3) presents some of the most often used kernels for Support Vector Machines.

$$\begin{aligned} K_{linear}(x, x') &= x \cdot x' \\ K_{polynomial}(x, x') &= (\gamma(x \cdot x') + r)^d \\ K_{rbf}(x, x') &= e^{-\gamma\|x-x'\|^2} \\ K_{sigmoid}(x, x') &= tanh(\gamma(x \cdot x') + r) \end{aligned} \tag{3}$$

4. Experimental setup and methodology

In this section, we describe the software libraries and the experimental and testing methods used for running our experiments.

4.1. **Software libraries and experimental methods.** All of our experiments are performed using the **scikit-learn** machine learning library [14]. For Support Vector Machines, this in turn uses the **libsvm** library [5], which is known to be a very powerful library for SVMs. Using well-known open source and well documented libraries ensures bug free experiments and guarantees that our experiments can easily be reproduced by other researchers.

We run experiments using the linear, RBF (Radial Basis Function), polynomial and sigmoid kernels. We use a randomized grid search to tune the hyperparameters for our SVM model (such as the C and ε values). It has recently been shown that using a random search is better than using a standard grid search [13]. We use 10 fold cross validation [6] as the model evaluation method within the random search. The parameter configuration that gives the best results, according to the Mean Absolute Error (MAE), is returned after 1000 random parameter samplings from a uniform distribution over $[0, 1)$ for each model parameter, except the d parameter for the polynomial kernel, which was sampled from the set $\{1, 2, 3, 4, 5\}$, and the C parameter, which was sampled from $[0, 10000)$ when optimizing the RBF kernel, from $[0, 10)$

when optimizing the sigmoid kernel and from $[0, 100)$ when optimizing the linear kernel. For the polynomial kernel, the grid searches found $d = 1$, which basically considers a linear kernel, so we do not include it in the presentation.

We also normalized our data by mean subtraction and division by the standard deviation.

For finding the optimal hyperparameters, we have added the normalization step as the first step of a pipeline, with the second and final step being the Support Vector Regressor. Our normalizer only scales features. The resulting pipeline is then used as the final regressor given to the randomized search for optimization. This entails that mean and standard deviation are computed on the training folds and the same values are used to normalize the test fold during testing. After normalization, the test fold is fed to the actual regressor part of the pipeline.

The way in which we optimize hyperparameters using randomized grid searching is fixed (we will refer to it as the **M1** method). However, we consider the following methods as well, which we will optimize using a standard grid search over a feasible set of hyperparameters.

(1) Method **M2** involves normalizing our entire data prior to doing 10 fold cross validation and also normalizing our targets (the correct statures). Performance scores are reported considering the values returned by the model unscaled. This approach can potentially overestimate the performance of a model, due to the fact that it does not really mimic real world scenarios by assuming we can also normalize the test data together with the training data;
(2) Method **M3** uses the pipeline approach of **M1**, but it also normalizes targets. We believe this to be a realistic application scenario that can help improve performance.

4.2. **Testing method.** For each methodology and kernel, we use a single run of 10 fold cross validation per iteration, storing and using the hyperparameters that define the model which minimizes the Mean Absolute Error score (MAE) over all iterations. That model is then used to report the MAE and Standard Error of the Estimate (SEE) according to Formula (4). In this formula, m is the number of considered instances, y_i is the known target value, and $\hat{y}_i$ is the value given by our trained model.

A 95% confidence interval for the mean on the 10 test folds is also reported, as described in [4].

We present our results for two case studies representing different populations as well as for their concatenation.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \qquad SEE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{4}$$

5. Data sets and experimental results

In this section we present the experimental evaluation of the SVM model on three case studies, following the process described in Section 4.

5.1. **Data sets.** Our data set is open source and taken from [2]. It consists of two case studies, both in the same format and containing seven features related to skeletal bone measurements in centimeters and the skeleton gender: the length of the humerus, the radius and ulna lengths, the femur, tibia and fibula lengths, the length of the whole leg (femur+tibia) and the gender. For each of the bones, the measurement represents the length of the longest of the two bones. Each of the two case studies contains 92 instances: the first one contains measurements of caucasians and the second one measurements of afro-americans (47 males and 45 females in each).

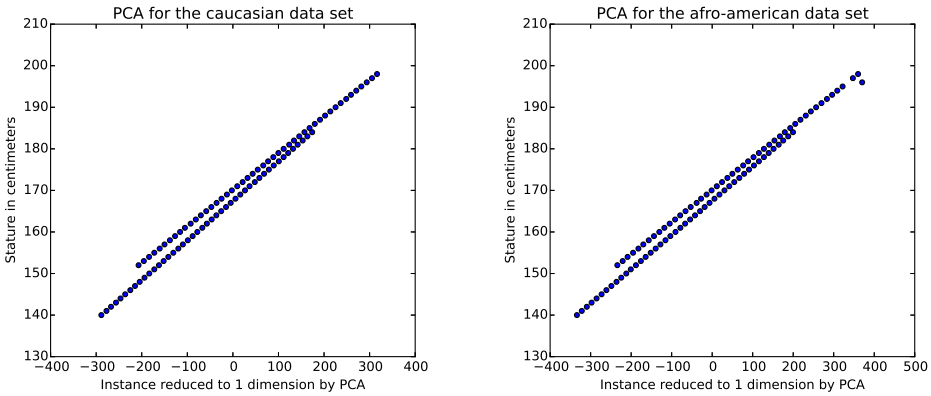
5.2. **Experimental results.** Figure 1 presents the two case studies reduced to a single dimension using Principal Component Analysis (PCA) [11]. The x axis of this graph represents the value of the single feature computed by PCA and the y axis represents heights. It can be seen that even under this setting, it is trivial to find a linear function that approximates the data, which makes us expect very good results from the linear SVM kernel, since it will be able to make use of more features, when even a single one shows good potential, at least if obtained by PCA.

5.2.1. *First case study - Caucasians.* Table 1 presents results for the first case study under all three evaluation methodologies.

For the Caucasian case study, the linear kernel proved to be the best under the **M1** testing methodology, while the RBF kernel proved to be the best under **M2** and **M3**.

For the **M2** and **M3** testing methodologies, we obtained the best results with identical model hyperparameters. This is understandable, since the methods are very similar and the scaled targets cannot differ too much in our data set. For larger datasets, the differences could potentially be bigger, thus requiring different hyperparameters. Therefore, one might want to employ a randomized grid search for the other methodologies as well. For our purposes however, we wanted to showcase both kinds of searches.

We note that all three kernels give very good results under all three methodologies, considering that the errors are in centimeters and the data



(A) Caucasian case study (B) Afro-american case study

FIGURE 1. Data set reduced to a single feature using Principal Component Analysis.

Kernel	**MAE (cm)**	**SEE (cm)**	**Hyperparameters**	**M**
linear	0.046±0.010	0.056±0.011	$C = 91.92, \varepsilon = 0.07$	M1
RBF	0.088±0.061	0.101±0.068	$C = 9625.77, \varepsilon = 0.07, \gamma = 0.004$	
sigmoid	0.881±0.504	0.913±0.504	$C = 4.064, \quad \varepsilon = 0.114,$ $\gamma = 0.0162, \quad r = 0.164$	
linear	0.049±0.009	0.057±0.009	$C = 5, \varepsilon = 0.001$	M2
RBF	0.042±0.014	0.050±0.015	$C = 910, \varepsilon = 0.0001, \gamma = 0.001$	
sigmoid	0.490±0.350	0.515±0.355	$C = 5, \quad \varepsilon = 0.0001,$ $\gamma = 0.01, \quad r = 0.0001$	
linear	0.049±0.010	0.057±0.010	$C = 5, \varepsilon = 0.001$	M3
RBF	0.040±0.012	0.048±0.014	$C = 910, \varepsilon = 0.0001, \gamma = 0.001$	
sigmoid	0.735±0.515	0.756±0.520	$C = 5, \quad \varepsilon = 0.0001,$ $\gamma = 0.01, \quad r = 0.0001$	

TABLE 1. Results obtained on the Caucasian case study. 95% confidence intervals are used for the results.

set contains almost 100 instances, which means that even in the worst case (0.881 MAE for the sigmoid kernel in the **M1** methodology), our average mistake is under one centimeter.

Overall, the best results are obtained under the **M3** methodology.

Kernel	MAE (cm)	SEE (cm)	Hyperparameters	M
linear	0.031±0.008	0.041±0.017	$C = 28.41, \varepsilon = 0.017$	M1
RBF	0.107±0.097	0.141±0.131	$C = 7304.752, \varepsilon = 0.037, \gamma = 0.02$	
sigmoid	0.268±0.099	0.339±0.171	$C = 9.844, \quad \varepsilon = 0.013,$ $\gamma = 0.0062, \quad r = 0.479$	
linear	0.056±0.052	0.114±0.154	$C = 0.1, \varepsilon = 0.001$	M2
RBF	0.051±0.030	0.081±0.072	$C = 910, \varepsilon = 0.001, \gamma = 0.001$	
sigmoid	0.146±0.090	0.187±0.141	$C = 0.9, \quad \varepsilon = 0.0001,$ $\gamma = 0.01, \quad r = 0.0001$	
linear	0.057±0.053	0.116±0.157	$C = 0.1, \varepsilon = 0.001$	M3
RBF	0.055±0.037	0.090±0.091	$C = 910, \varepsilon = 0.001, \gamma = 0.001$	
sigmoid	0.195±0.147	0.220±0.167	$C = 0.9, \quad \varepsilon = 0.0001,$ $\gamma = 0.01, \quad r = 0.0001$	

TABLE 2. Results obtained on the African-american case study. 95% confidence intervals are used for the results.

5.2.2. *Second case study - Afro-americans.* Table 2 presents results for the second case study under all three evaluation methodologies.

Once more, the linear kernel is the best under the **M1** methodology. Compared to the first case study's **M1** results, scores are better with the linear and sigmoid kernels and worse with the RBF kernel, although the differences are very small for any practical concerns.

The RBF kernel took the top spot under the **M2** and **M3** methods once again. Compared with the first case study, the **M2** and **M3** results are worse for the Afro-american case study, only the sigmoid kernel in the **M2** setting managing to surpass its direct competitor.

Overall, results are worse on the second case study than on the first, but not in a significant fashion.

This time, the best results are obtained under the **M1** methodology.

5.2.3. *Mixed case study.* The mixed case study consists of the concatenation of the data sets corresponding to the previous two case studies, resulting in a bigger data set that contains both populations.

We have not added any new feature to distinguish the two populations.

Since the two populations look similar in the PCA plots (Figure 1), we expect their concatenation to yield good results as well.

Table 3 presents results for the mixed study. This time, the radial basis function kernel clearly outperforms the other kernels taken into consideration.

Kernel	MAE (cm)	SEE (cm)	Hyperparameters	M
linear	0.340±0.103	0.417±0.129	$C = 98.508, \varepsilon = 0.394$	M1
RBF	0.243±0.152	0.347±0.223	$C = 9563.38, \varepsilon = 0.164, \gamma = 0.009$	
sigmoid	2.604±1.177	2.855±1.208	$C = 5.824, \quad \varepsilon = 0.0972,$ $\gamma = 0.01, \quad r = 0.7156$	
linear	0.458±0.225	0.551±0.254	$C = 0.29, \varepsilon = 0.001$	M2
RBF	0.261±0.137	0.367±0.215	$C = 10000, \varepsilon = 0.0001, \gamma = 0.001$	
sigmoid	1.246±0.391	1.305±0.389	$C = 6, \quad \varepsilon = 0.1,$ $\gamma = 0.01, \quad r = 0.0001$	
linear	0.469±0.223	0.577±0.270	$C = 0.29, \varepsilon = 0.001$	M3
RBF	0.252±0.130	0.376±0.268	$C = 10000, \varepsilon = 0.0001, \gamma = 0.001$	
sigmoid	1.469±0.498	1.585±0.527	$C = 6, \quad \varepsilon = 0.1,$ $\gamma = 0.01, \quad r = 0.0001$	

TABLE 3. Results obtained on the mixed case study. 95% confidence intervals are used for the results.

This implies that the RBF kernel is the most robust, being able to deal with more data instances even if they are from different populations. This suggests that the RBF kernel should perform the best in practice.

For the **M2** and **M3** methodologies, we again did not obtain significant differences in results with different model hyperparameters.

We obtained the best results under the **M1** methodology.

Another testing scenario that we plan to research in the future involves the concatenation of the two data sets, but with a new feature added that specifies which population each instance belongs to.

Figure 2 shows the learning curves for the RBF kernel on the mixed case study, unde the **M1** testing methodology. It can be seen the model generalizes very well and there is no overfitting. Because of the good generalization, it is unlikely for more data to produce better results. Because the training score error increases very slowly, it is likely that more data will not lead to worse results. It can also be seen that the model achieves its optimal performance with few training instances.

6. Comparison to related work

All of our results outperform existing literature results. As far as we are aware of, only [2] presents results on these data sets. Their obtained SEE

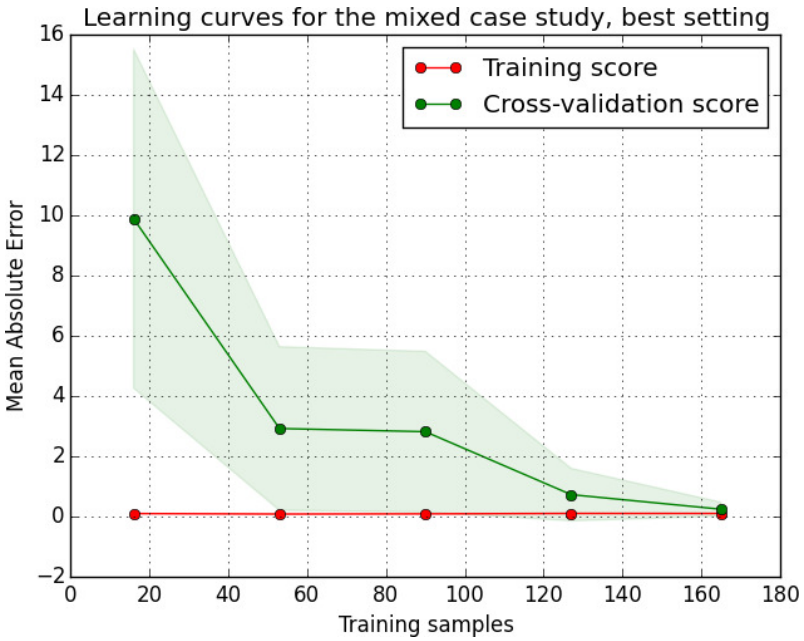


FIGURE 2. Learning curves for the mixed case study, with the RBF kernel under the **M1** methodology and considering MAE scores.

errors are between 3.05 and 3.66 cm. Our worst result is 2.855 SEE on the mixed case study using the sigmoid kernel under the **M1** testing methodology.

On the individual case studies, all of our results are well below 1 SEE, with most of them being under 0.5 and the best of them under 0.1.

Taking into consideration work done on other data sets but regarding the same issue, such as [17, 3, 12, 19] and others, we note that no previously existing method manages to achieve errors less than a centimeter, under any scoring metric. This can be attributed to the fact that previous methods only seem to consider a few features, to which they apply rather basic statistical procedures. It is impossible for us to test SVM applications on those data sets, since they are not public. Therefore, a direct comparison between our results and existing results on the data sets in [17, 3, 12, 19] would be meaningless. However, it stands to reason that, given the results we have presented in Section 5, SVM-based models can potentially outperform existing methods on other data sets as well.

Table 4 presents a quick comparison between our methods and other literature results on the data set we have worked with.

Method	Best SEE	Worst SEE	Short description
SVM	0.048	0.913	First case study, RBF kernel under the **M3** methodology for the best SEE and sigmoid kernel under the **M1** methodology for the worst case.
SVM	0.041	0.339	Second case study, linear and sigmoid kernels, both under the **M1** methodology.
SVM	0.347	2.855	Mixed case study, RBF kernel under the **M1** methodology and sigmoid kernel also under the **M1** methodology.
[2]	3.05	3.66	Regression formulas based on basic statistical methods.

TABLE 4. Overview of literature results on the data set we have used.

Therefore, our Support Vector Regression approach is a very reliable solution to the problem of skeletal height estimation given the lengths of certain bones, leading to much better results than previous approaches and having the potential to be easily adapted to new data sets from the field.

7. Conclusions and future work

We have presented in this paper how SVM regression methods can be applied for estimating the height of archaeological remains. We have run extensive experiments on two archaeological data sets that are freely available, obtaining very good results that surpass previous results from the literature.

Our results are also superior considering other data sets. While this comparison does require further investigation, it is a valid conjecture due to the fact that SVMs are a family of machine learning algorithms, which means they can learn from any type of data. If they could learn well on one data sets, it stands to reason that they are able to do the same on another data set containing similar kind of data.

We have applied three testing methodologies, which we believe to mimic certain real world scenarios well. We also used a randomized grid search for one of them, which recent research has shown to be the best was of optimizing hyperparameters.

Therefore, we consider the SVM-based methods that we have applied to offer significant contributions to the fields of archaeology and forensic analysis and believe that they will generalize well to other problems of a similar nature.

As a future research direction, we plan to experiment with more machine learning libraries and on more data sets. We also plan to find different kernels to test. Another possible direction is refining our hyperparameter searches, by reducing the intervals we search in and by running the randomized search for more iterations, increasing the probability of finding better hyperparameters.

If we can obtain more data, we believe that researching online learning options would also be a useful undertaking.

Since the PCA reduction did not seem to generate useless data, it is also worth investigating the results that can be obtained using less features, since fewer measurements should always be helpful in practice.

References

[1] A. Smola., B. Schlkopf, *A tutorial on support vector regression*, Statistics and Computing, 14 (2004), no. 3, pp. 199-222.

[2] M. Trotter, G. Gleser, *Estimation of Stature from Long Bones of American Whites and Negroes*, American Journal of Physical Anthropology, **10**(1952), no. 4, pp. 463-514.

[3] D.C. Pal, A.K. Datta, *Estimation of stature from radius length in living adult Bengali males*, Indian Journal of Basic and Applied Medical Research, 3 (2014), pp. 380-389.

[4] L. Brown, T. Cat, A. DasGupta, *Interval estimation for a proportion*, Statistical Science, 16 (2001), pp. 101-133.

[5] C.-C Chang, C.-J., Lin, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2 (2011), no. 3, pp. 27:1-27:27.

[6] T.M. Mitchell, *Machine Learning*, McGraw-Hill, Inc. New York, USA, 1997.

[7] G. Czibula, V. Ionescu, D. Miholca, I. Mircea, *Novel supervised learning based approaches for stature prediction of archaeological skeletal remains from bones*, Journal of archaeological science, under review.

[8] P.H. Stevenson, *On racial differences in stature long bone regression formulae, with special reference to stature reconstruction formulae for the chinese*, Biometrika Trust, 21 (1929), pp. 303-318.

[9] A. Telkka, *On the Prediction of Human Stature from the Long Bones*, Acta Anatomica, 9 (1950), no. 1-2, pp. 103-117.

[10] C.W. Depertuis, J.A. Hadden, *On the reconstruction of stature from long bones*, American Journal of Physical Anthropology, 9 (2005), no. 1, pp. 15-54.

[11] M.E. Tipping, C.M. Bishop *Probabilistic principal component analysis*, Journal of the Royal Statistical Society, Series B, 61 (1999), pp. 611-622.

[12] M.R. Feldesman, *Race Specificity and the Femur/Stature Ratio*, American Journal of Physical Anthropology, 100 (1996), no. 2, pp. 207-224.

[13] J. Bergstra, Y. Bengio, *Random search for hyper-parameter optimization*, Journal of Machine Learning Research, 13 (2012), pp. 281-305.

[14] F. Pedregosa et al., *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825-2830.

[15] N. Navsa, *Skeletal morphology of the human hand as applied in forensic anthropology*, Ph.D. thesis, University of Pretoria, 2010.

[16] C. Cortes, V. Vapnik, *Support-vector networks*, Machine Learning, 20 (1995), no. 3, pp. 273-297.

[17] D. McCarthy, *The long and short of it: The reliability and inter-populational applicability of stature regression equations*, Master's thesis, Oregon State University, 2001.
[18] J.K. Lundy, *The mathematical versus anatomical methods of stature estimate from long bones*, The American Journal of Forensic Medicine and Pathology, 6 (1985), pp. 73-75.
[19] G. Fully, *Une Nouvelle Methode de Determination De la Taille*, Annales de Medecine et de Criminologie, 36 (1956), pp. 266-273.

"Babeş-Bolyai" University,, Faculty of Mathematics and Computer Sciences, 1, Kogălniceanu Street,, 400084 Cluj-Napoca,, Romania
E-mail address: ivlad@cs.ubbcluj.ro