Doi:10.47572/muskutd.763218

Estimation of Bone Age from Radiological Images with Machine Learning

Makine Öğrenmesi ile Radyolojik Görüntülerden Kemik Yaşı Tahmini

Nida GÖKÇE NARİN[1], İbrahim Önder YENİÇERİ[2], Gamze YÜKSEL[3]

[1]Department of Statistics, Faculty of Science, Mugla Sıtkı Kocman University, Muğla
[2]Department of Radiology, Faculty of Medicine, Mugla Sıtkı Kocman University, Muğla
[3] Department of Mathematics, Faculty of Science, Mugla Sıtkı Kocman University, Muğla

Öz

Kemik yaşı tahmini, endokrinolojik sorunların ve adli sorunların tanısında önemlidir. Greulich ve Pyle (GP) yöntemi kemik yaşı tahmini için yaygın olarak kullanılmaktadır. Ancak, gözlemcinin kendisi ve gözlemciler arası nispeten yüksek bir değişkenliğe sahiptir. Bu nedenle, kemik yaşının hesaplanmasında uzmanlardan bağımsız otomasyon tabanlı sistemler geliştirilmeye başlanmıştır. Bu çalışmada, makine öğrenimine dayalı sınıflandırma yöntemlerinin kemik yaşı tahmin performanslarını karşılaştırmayı amaçladık. Çalışmaya 12-108 aylık 388 erkek ve 387 kız dahil edildi. Cohort el bilek grafilerinde kemik alanın tüm el bilek alanına oranı her olgu için hesaplandı ve olgular üçer aylık intervaller ile sınıflandırıldı. Bu, veri tabanı olarak kabul edilip test verisi bu veri tabanı ile test edildi. Kemik yaşı tahmini için makine öğrenmesine (ML) dayanan tahmin modellerini kullandık. Weka ara yüzü kullanılarak oluşturulan modellerin tahmini performansları kronolojik yaş ile karşılaştırıldı. Ayrıca yöntemlerin öngörücü performansı arasında istatistiksel olarak anlamlı bir fark olup olmadığı Friedman testi ile test edilmiştir. Sonuç olarak, kız çocukları için ML yöntemleriyle yapılan kemik yaşı tahmininin kronolojik yaş ile anlamlı derecede ilişkili olduğu gözlenmiştir. GP ve kronolojik yaş arasında anlamlı bir fark bulundu. Bu çalışmadan elde edilen sonuçlar, ML tabanlı sınıflandırma yöntemlerinin kemik yaşını tahmin etmede yüksek başarı gösterdiğini göstermiştir. Bu nedenle, ML sınıflandırma modellerinin kemik yaşını tahmin etmek için kullanılabileceğini önermekteyiz.

Anahtar Kelimeler: Bilek Radyografisi, Kemik Yaşı Tahmini, Makine Öğrenmesi

Abstract

Bone age estimation (BAE) is important in the diagnosis of endocrinological problems and forensic issues. Greulich and Pyle (GP) method is widely used for BAE. But it has relatively high intraobserver and interobserver variability. For this reason, automation-based systems independent of experts have started to be developed in estimating bone age. We aimed to compare bone age estimation performances of machine learning based classification methods. A total of 388 boys and 387 girls between the age of 12-108 months were included in the study. In Cohort wrist radiographs, the ratio of bone area to the entire wrist area was calculated for each case, and the cases were classified with quarterly intervals. This is considered as a database and the test data has been tested with this database. We used the estimation models which are based on Machine learning (ML) for BAE. The predicted performances of the models created by using Weka interface were compared with chronological age. Moreover, whether there is a statistically significant difference between the predictive performance of the methods was tested by the Friedman test. As a result, it was observed that bone age estimation performed with ML methods for girls was significantly correlative with chronological age. A significant difference was found between GP and chronological age. The results obtained from this study showed that ML-based classification methods have high success in predicting bone age. Therefore, we suggest that ML classification models can be used to predict bone age.

Keywords: Bone Age Estimation, Machine Learning, Wrist Radiography

Introduction

Bone age estimation (BAE) is important in the diagnosis of endocrinological problems and forensic issues. Radiographs of different skeletal regions are frequently used in bone age estimation. Studies have shown that wrist radiographs in children under the age of 16 are in good agreement with the age of the bone. Therefore, wrist radiographs are frequently used in age prediction of children under the age of 16 years. The ossification points in the wrist region follow a sequence that is usually stable during development (1). In the wrist region, there are 11 ossification regions, 8 of which are carpal bones and 3 are epiphyseal. The first ossification in the wrist region starts with capitated bone, followed by hamate (Ham), radial head epiphysis, triquetrum (Trq), 1st metacarpal epiphysis (1.MC), lunate (Lnt), trapezium (Trzm), trapezoid (Trzd), ulna epiphysis, scaphoid (Scph), and pisiforme (Ps) bones.

There are numerous studies in the literature predicting bone age according to the radiographs of the wrist or the different bone regions in the body. Greulich and Pyle (GP) atlas and Tanner and Whitehouse (TW) are commonly used BAE methods (1-6). The GP method uses an atlas that has been previously obtained from radiographs, which is standardized for all ages. The radiographs of the patient to be evaluated are compared to the age group in this atlas and estimated to be the nearest group. Due to its simplicity and speed, this method has become the most widely used reference standard. However, it has relatively high intraobserver and interobserver variability (7). In the TW method, several specific regions of interest (ROI) in the hand are assessed. A numerical score is associated with each stage of each bone. By adding the scores of all ROIs, the overall maturity score is obtained.

	ORCID No
Nida GÖKÇE NARİN	0000-0002-4840-5408
İbrahim Önder YENİÇERİ	0000-0003-2779-2020
Gamze YÜKSEL	0000-0003-3578-2762

Başvuru Tarihi / Received: 03.07.2020
Kabul Tarihi / Accepted : 01.02.2021

Adres / Correspondence : İbrahim Önder YENİÇERİ
Department of Radiology, Faculty of Medicine, Mugla Sıtkı Kocman University, Muğla
e-posta / e-mail : onderyeniceri@mu.edu.tr

Due to the complexity of the calculation, it is not as widely used as GP (8).

In recent years, new methods of BAE are studied. The support vector machines and fuzzy methods are used for bone age estimation of the pediatric population (9-11). However, there are many machine learning methods which are developed with increasing the power of computers. Neural Networks are widely used for the classification of pediatric X-rays images (12-15). Machine Learning methods are also presented in the literature (16,17). In recent years, studies on the prediction of bone age have attracted the attention of software developers and they tried to predict bone age more accurately with convolutional neural networks and deep learning techniques (18-21). Apart from these, BoneXpert and 16 bit have recently been put into clinical use in software that estimates automatic BAE (22,23).

In this study, we aimed to compare bone age estimation performances of machine learning based classification methods. The bone age estimation performances of the classification models developed by using boys' and girls' cases with known chronological age (CA) were compared with both CA and GP.

Material and Method

Study population

This research is a retrospective study. Ethics committee permission was obtained for this study with the protocol number 5514 of Muğla Sıtkı Koçman University Human Research Ethics Committee dated 13.03.2017. Data were obtained by using the old records in the hospital PACS system (PACS; Sisoft). The study population consisted of wrist radiographs taken from the patients who came to the radiology department of our institution. Cases with fracture and deformation in the wrist region, skeletal dysplasia or metabolic disease anamnesis were not included in the study. Since eight of the male cases were diagnosed with a developmental disorder, these cases were not included in the training set during the machine learning process. As a result of a preliminary evaluation of descriptive statistics, the cases which were determined as outliers were excluded from the data set. In this study, the cases with a missing birth date were not included in the study because the success of the classification methods will be evaluated by considering the CA. A total of 388 boys and 387 girls between the age of 12-108 months were included in the study.

Radiological data and measurements

Wrist radiographs of the cases were recorded as DICOM files in the archive of our hospital. The images were examined with the SISOFT DICOM viewer on the medical monitor (Sisoftdicom viewer, Ankara/TURKEY). Digital images of the cases were evaluated by a 20-year radiologist working in the field of general radiology.

A radiography of bone development in a wrist region with a completed individual is given in Fig. 1. The layout of the bones is seen in the radiograph. There are joint spaces between the bones. On the 2D radiograph, triquetrum and pisiform and trapezium and trapezoid seem to overlap. However, there are also joints between these bones.

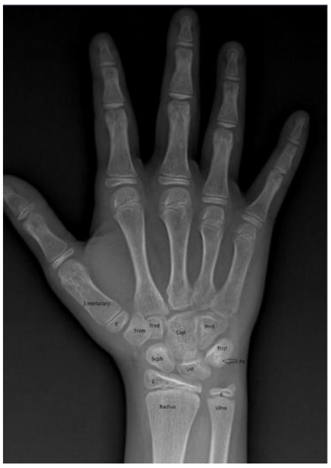

Figure 1. Left Hand Wrist Graph

In this study, the rate of ossification (RO) was obtained by dividing the area of the total ossification points (TOP) in the carpal region by the area of the carpal region (CR). This ratio is included in the data set as a variable. The ossification points in the carpal region were measured using measurement tools in the PACS program. Fig. 2 shows how these areas are calculated. Fig. 2a shows how to calculate the TOP value of a 15-month-old male case. The total area was calculated by drawing each bone outer contour with the PACS program. Fig. 2b shows how the CR is calculated for the same situation. As the age increases, the bones are superposed on each other due to the increase in the maturity of the bones. Therefore, the contours of the superposed bones were measured over the other bone to ensure that the actual size of each bone was obtained. Fig. 3a and 3b show the CR and TOP measurements in a 96-month-old girl, respectively. Note that trapezium-trapezoid and triquetrum-pisiform measurements are superimposed on each other in Fig. 3a.

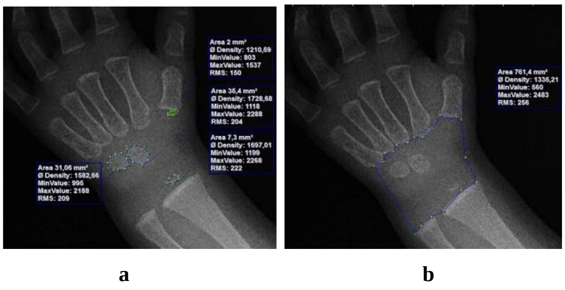

Figure 2. a: The total area of the ossification point (TOP) in carpal region. b**:** The area of the carpal region (CR)

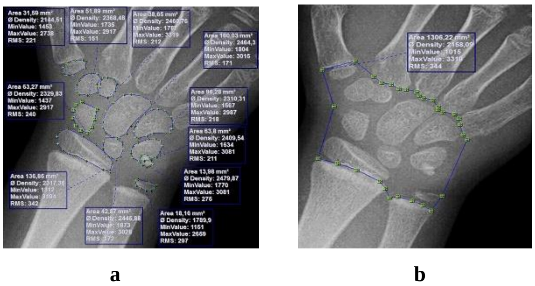

Figure 3. a: For a 96-month-old girl the TOP value calculation example. **b:** For a 96-month-old girl the CR value calculation example

Machine learning based classification

One of the problems that ML methods are commonly used is the classification problem. ML methods are successful in solving classification problems that aim to model the relationship between independent variables and categorically dependent variables. Many algorithms make classification modeling based on machine learning approaches such as artificial neural networks, decision tree, Bayes classifier, and logistic regression (24-28).

An ML model learns how each of the data properties called variables is associated with different outputs. Many complex problems for estimating bone age can be solved by designing the correct properties of the problem and then modeling these properties with a simple ML algorithm.

In this study, ML-based Multilayer Perceptron, Bayesian Networks, Multinomial Logistic Regression, Logistic Model Tree prediction models were used to estimate bone age. The models were created using the Weka interface (29). Estimated performances of the models were determined by comparing with chronological age. Besides, the Friedman test was used to test whether there is a statistically significant difference between the predictive performance of the methods (30-31).

Multilayer Perceptron (MLP)

Multi-Layer Perceptron (MLP) is a classifier with a feed-forward neural network architecture that maps a set of input values to output values (25). It is used in the solution of nonlinear problems. MLP generally uses the Delta learning rule to update the weights in the training process and the Gradient Descent algorithm for the optimization of the loss function. MLP architecture consists of an input layer, one or more hidden layers, and an exit layer. Each layer can have a different number of neurons, and each layer can be completely connected to the next. Network architecture, learning rule, and selection of optimization algorithms may vary depending on the problem in question. During the learning phase, the network learns by adjusting the weights to guess the correct class label of the input data (25).

Bayesian Networks

Bayesian Networks (BN) is a classifier based on Bayes' theorem. They are represented by directional acyclic graphs. Bayes Theorem is given as follows: P(A│B)=(P(B|A)P(A))/(P(B)), P(B)>0$P(A)$ and $P(B)$ are marginal probability of events A and B respectively. $P(A|B)$ is the conditional probability of A given B, $P(B|A)$ is the conditional probability of B given A.

In BN, edges represent conditional dependencies, and nodes represent a unique random variable. They are used to model complex systems. Its purpose is to model the distribution probabilities of variables, conditional dependency, and causality by using the observation of some of the independent variables and the prior knowledge of others. (27).

The multinomial logistic regression

The multinomial logistic regression (MLR) model is a generalized form of the binary logistic regression model that includes more than two categories and is used to model different selections (28). Logistic regression analysis is a method for explaining cause-and-effect relationships between dependent variables and independent variables. Although it is called regression, logistic regression is a classification method where the dependent variable is categorical.

Decision Trees

Decision Trees (DT) is a classification method that aims to divide a data set containing a large number of observations into smaller sets using a set of rules. It consists of branches starting from a root node and descending downward. Both categorical and numerical data can be used in the classification. The DT consists of 3 main components: the root node, the inner node and the leaf node (33). Inner nodes represent a state based on the division of the tree into branches/edges. Leaf nodes represent a decision. In real data sets with many features, DT can produce simple and fast solutions. DT makes variable selection or property selection. A significant advantage of DT is that nonlinear relationships between variables do not affect tree performance.

The Logistic Model Tree

The logistics model tree (LMT) is a machine learning method obtained by combining logistic regression and decision tree. It is a standard decision tree structure with logistic regression functions on leaves. Logit Boost algorithm is used to create a logistic regression model from each node of the tree (32,33).

Machine learning needs to confirm the stability of the models. The model obtained from the training data must ensure that it will make the correct prediction for the actual data. That is, the model should assure that bias and variance are low for data that does not contain much noise. The simplest validation technique is known as the Holdout method. It divides the data set into two groups as training and testing. It evaluates the performance of

Doi:10.47572/muskutd.763218

the model obtained from the training set using the test set. However, this method has a high variance problem. The k-Fold Cross Validation method has solved this problem by dividing the data into k sub-groups and applying the Holdout method k times. This method significantly reduces variance and bias. As a result of experimental studies, the value of k usually takes values of 5-10 but is not mandatory.

Performance Criteria

There are different criteria for measuring the estimated performance of a classification model. The most commonly used of these are the accuracy, precision and recall calculated with the confusion matrix given in Table 1. Accuracy (ACC) is a measure of the correct estimation rate of the classifiers. Precision (PRE) is the ratio of the number of positively classified positive observations of the total number of positive predicted observations. Recall (REC) is the ratio of the number of positively classified positive observations of the total number of positive real observations (34).

Table 1. Confusion Matrix

		Predicted Class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

True Positive (TP) is that the model predicts the positive class correctly; *False Negative (FN)* is the result that the model's negative class incorrectly predicts*;* True Negative (TN) is that the model predicts the negative class correctly; False Positive (FP) is a result in which the model predicts the positive class incorrectly.

Statistics

Descriptive statistics regarding the data set used in this study are given in Table 2. The data set was classified using five different machine learning methods with the Weka program. For the classification, the Bayesian classifiers Naive Bayes and Bayes net, LMT which is one of the decision tree classifiers, multi-layer perceptron, which is a classifier based on artificial neural networks, and multinomial logistic regression methods were used. The data set was trained separately for both girls and boys using the 10-fold cross-validation approach.

Table 2. Descriptive statistics of data set

	Boys				Girls			
Months	n	Mean	Std. Dev.	Std. Error	n	Mean	Std. Dev.	Std. Error
12	10	7.86	2.87	0.91	12	8.54	1.92	0.56
15	11	7.93	1.72	0.52	12	11.53	3.02	0.87
18	13	10.20	3.50	0.97	12	12.27	4.95	1.43
21	9	12.51	2.43	0.81	12	14.11	3.28	0.95
24	11	12.08	3.82	1.15	11	15.67	4.67	1.41
27	11	14.28	2.63	0.79	12	19.17	6.86	1.98
30	15	13.92	3.94	1.02	11	19.87	4.48	1.35
33	13	16.60	2.10	0.58	12	18.86	3.48	1.00
36	11	16.94	5.27	1.59	13	24.43	7.61	2.11
39	12	17.17	3.89	1.12	13	24.20	6.29	1.74
42	11	19.99	4.91	1.48	10	29.29	6.29	1.99
45	12	21.24	5.26	1.52	11	25.33	4.80	1.45
48	12	23.08	8.22	2.37	10	24.86	3.51	1.11
51	11	22.06	4.26	1.29	12	27.37	9.09	2.63
54	12	22.73	6.29	1.82	11	34.61	8.20	2.47
57	13	26.68	4.71	1.31	10	37.75	5.96	1.88
60	14	26.33	3.65	0.97	12	42.31	9.85	2.84
63	10	27.96	5.10	1.61	13	39.93	6.23	1.73
66	11	30.49	6.06	1.83	13	43.30	7.91	2.19
69	12	30.05	4.22	1.22	13	43.94	5.53	1.53
72	11	30.98	5.78	1.74	12	44.40	5.06	1.46
75	10	34.09	6.90	2.18	12	48.37	8.41	2.43
78	10	35.71	10.13	3.20	11	50.20	8.21	2.48
81	13	38.12	7.82	2.17	12	53.25	8.62	2.49
84	14	40.61	7.11	1.90	12	58.02	6.89	1.99
87	13	47.93	8.97	2.49	10	55.39	5.71	1.80
90	12	41.33	8.68	2.50	11	61.95	11.10	3.35
93	12	44.57	10.27	2.96	12	63.83	6.28	1.81
96	11	49.51	9.89	2.98	12	64.45	8.16	2.36
99	13	50.86	6.05	1.68	11	64.86	11.49	3.47
102	12	50.68	8.08	2.33	11	66.59	9.33	2.81
105	11	55.64	10.01	3.02	13	69.85	5.15	1.43
108	12	61.00	8.50	2.45	13	74.91	6.61	1.83
Total	**388**				**387**			

n represents the number of girls and boys for each month, Mean is arithmetic mean of RO variable, Std.Dev. is standard deviation of RO, Std. Error is standard error of RO in Table 2.When mean values are examined, it is seen that RO values are very close to each other in close months. Even for some month groups, RO values were obtained smaller than the previous month values. This difference arises from the cases under consideration. Also, it is seen that the bone development of girls is faster than boys for the same month groups.There is no significant difference between standard deviation and standard error values by gender.

The optimal values of the training parameters were determined by trial and error. For all algorithms, the iteration number is 1000 and the batch size is 100. Simple estimator and K2 search algorithms were used for classification with Bayes Net. In the classification with multilayer perceptrons, the learning rate is 0.3, momentum is 0.2, training time is 500, and validation threshold 20 were selected. A ridge estimator was used in multinomial logistic regression. In classification with LMT, fast regression is true, the number of boosting iterations is -1, and weightTrimBeta is 0 were taken. Finally, Friedman test was used to test whether there is a significant difference between the prediction performances of ML-based classification models.

Results

Descriptive statistics for the study population are given in Table 2. "n" represents the number of girls and boys for each month, Mean is arithmetic mean of the RO variable, Std. Dev. is the standard deviation of RO, Std. Error is the standard error of RO in Table 2. When the average values are examined, it is seen that the RO values are very close to each other in recent months. Even for some month groups, RO values were obtained lower than the previous month values. This difference is due to the cases examined. Also, for the same month groups, it was observed that the bone development of girls was faster than boys. There was no significant difference between standard deviation and standard error values according to gender according to the Mann-Witney-U test (pstandDev=0.296>0.05; pstandErr=0.265>0.05).

The comparison of the performance of GP and ML and classification methods with CA are given in Table 3. In Table 3, TPR is the ratio of the true classified cases to the total number of cases. Rec is the ratio of the number of positive observations, which are classified as positive, by the total number of positive observations. Pre is the ratio of the number of positive observations positively positive to the number of positive observations. Acc is a measure of the correct estimate rate of classifiers. All values in the table are expected to be close to 1, which is the best performance. ± 3M, ± 6M, ± 9M, and ± 12M represent the correct classification rate for all estimation methods with 3, 6, 9, 12-month deviations, respectively. As a result, it was seen that the best prediction performance was obtained with the MLR method in boys, while the LMT method was better in girls.

Table 3. Comparison of ML classification methods performances vs GP and CA

	Methods	TPR	Rec	Pre	Acc	±3M	±6M	±9M	±12M
Boys	**GP**	0.12	0.17	0.12	0.92	0.31	0.41	0.54	0.63
	BN	0.24	0.25	0.25	0.94	0.49	0.62	0.71	0.80
	LMT	0.28	0.28	0.30	0.94	0.46	0.59	0.71	0.79
	MLR	**0.38**	0.38	0.41	0.94	**0.53**	**0.67**	**0.79**	**0.85**
	MLP	0.37	0.35	0.36	0.97	0.56	0.67	0.75	0.82
Girls	**GP**	0.13	0.18	0.13	0.95	0.30	0.45	0.58	0.68
	BN	0.23	0.24	0.24	0.95	0.43	0.55	0.72	0.82
	LMT	0.72	0.71	0.71	0.98	0.79	0.85	0.91	0.95
	MLR	0.37	0.38	0.37	0.96	0.53	0.65	0.76	0.82
	MLP	0.35	0.79	0.37	0.96	0.49	0.61	0.71	0.83

TPR is the ratio of the correctly classified cases to the total number of cases. Accuracy (ACC) is a measure of the correct estimation rate of the classifiers. Precision (PRE) is the ratio of the number of positively classified positive observations of the total number of positive predicted observations. Recall (REC) is the ratio of the number of positively classified positive observations of the total number of positive real observations. All values in the Table 3 are expected to be close to 1, which is the perfect agreement. ±3M, ±6M, ±9M, and ±12M represent the ratio of 3, 6, 9, 12 monthly deviations between CA and ML-based estimation respectively.

Whether there was a significant difference between bone age prediction performances of the classification models was tested with Friedman test at the 0.05 significance level. The Friedman test is a nonparametric test that compares three or more paired groups. Friedman statistic is calculated from the sum of ranks and the sample sizes. When the sum of ranks for groups is very different from each other, the p-value is close to zero.

Table 4 shows the median and (25th and 75th) percentils values for bone age estimation of ML methods. As a result of the Friedman Test, a statistically significant difference (p <0.00001) was found among the classification methods' performances. Then, multiple comparison tests are performed to see which methods are different from others. Fig. 4 and Fig. 5 show the results of multiple comparison tests of the median for boy and girl cases, respectively. By default, the median of CA is highlighted and the comparison of the range is shown in blue. For the other groups, the intersections between the intervals of the CA averages are highlighted in red in those that do not intersect the gray. The intersection of the group median means that the methods are not significantly different from each other, otherwise, the group median is significantly different from each other.

In Table 5, the lower and upper limits of the 95% confidence intervals for the difference between the means of the groups compared from the multiple comparison test result and for the true mean difference are given. In the table, the first two columns represent the compared methods, the third column estimates the difference between the group

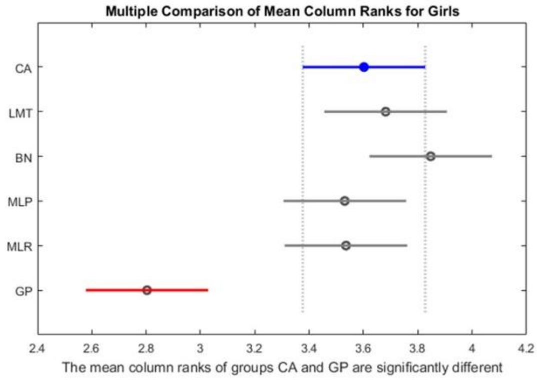

Figure 4. The result of multiple comparison test for girls

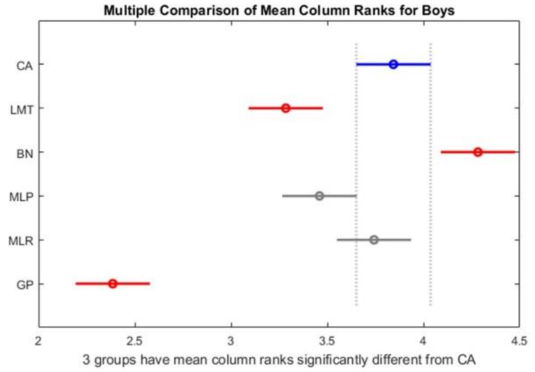

Figure 5. The result of multiple comparison test for boys

means (DEGM), and the last column represents the p-value for the hypothesis test where the corresponding mean difference equals zero. Small p values (p-values <0.05) indicate a significant difference between methods.

As a result, it was observed that bone age estimation performed with ML methods for girls was significantly compatible with CA. A significant difference was found between GP and CA. While the bone age estimation performed only with MLP and MLR among the ML methods for boys was found to be compatible with CA, the predictions made with LMT, BN and GP were found to be significantly different from CA.

Discussion

In this study, bone age in pediatric cases was estimated by machine learning approaches. Bone age assessment is very important for early diagnosis of growth disorders in pediatric cases and determining calendar age in forensic cases. Therefore, there is a need to develop high performance and stable methods for bone age estimation. The most common method used to estimate bone age is to use a standardized atlas for all ages (1). A radiologist compares the patient's radiographs with the age group in this atlas and estimates the group they finds closest. However, the accuracy rate of the estimate in this method may vary from expert to expert. Research has shown that experienced experts have higher predictive success (7,8).

In GP atlas bone age can be estimated at 2-4 months intervals under 3 years old, 6 months intervals between 3-6 years old, and 1-year intervals above 6 years old. In this study, cases between 12 and 108 months were selected for all groups at three-month intervals. Measurements were made with an accuracy of ± 15 days and no cases were used in the months in between.

In the estimations using GP atlas, all hand and wrist bones are used. In this study, only bones in the wrist (carpal) region were used for age estimation. Making a more consistent estimate with less information is an important advantage.

ML methods used in this study are the most preferred methods to make estimates based on classification when the dependent variable is a multi-class categorical variable. Bayesian classifiers make a classification based on conditional probability. Decision trees try to model the relationships between variables with decision rules. Artificial neural networks aim to model the relationship between the dependent variable and the independent variables with the black box principle. Multinomial logistic regression is a machine learning approach based on statistical models. In this study, the performances of these 4 different approaches, all of which can be used for the same purpose, on the prediction of bone age were examined.

In this study, ML methods were used to estimate bone age. The bone age estimation performed with MLP and MLR methods was found to be compatible with CA for both girls and boys, whereas GP tended to predict bone age less than it was. In addition, there is a statistically significant difference between GP and CA. Koç et al. in their study with Turkish children in the working group of 7-13 age groups underestimate the GP has found that the age underestimates (35). Our study consists of Turkish children in the 1-9 age group and likewise made the GP underestimate. Our results are consistent with the fact that the GP method has prepubertal estimates for most populations.

The harmony of estimation results with CA with ML methods was found higher than GP. This is because the data set generated from the study comes from a different population than the data set that makes up the GP atlas. Machine learning methods can make the most appropriate estimates for the population from which the data comes from since they learn the relationships from the data. The compliance of the obtained results with CA also confirms this skill. Moreover, the predictive performance of ML methods can be further improved by increasing the number of samples.

Since the GP method is an approach based on expert opinion, prediction results may differ from expert to expert, while machine learning methods give more stable results. Therefore, there is a need to develop automation systems based on machine learning for bone age estimation.

Table 4. Median (25th - 75th percentile) values for estimation of bone age

		Minimum	25th Percentile	Median	75th Percentile	Maximum
Boys	**CA**	12	36	60	86	108
	LMT	12	30	57	84	108
	BN	15	33	60	92	108
	MLP	12	33	57	84	108
	MLR	12	33	60	84	108
	GP	12	30	54	72	144
Girls	**CA**	12	39	63	84	108
	LMT	12	39	63	84	108
	BN	12	42	70	87	108
	MLP	12	42	63	84	108
	MLR	12	39	63	84	108
	GP	8	30	60	84	126

Table 5. Multiple comparison results

				95.0% Bounds		
	Group 1	Group 2	D_{EGM}	Lower	Upper	p-value
Boys	**CA**	**LMT**	0.5595	0.1737	0.9453	**0.0005**
	CA	**BN**	-0.4389	-0.8247	-0.0531	**0.0151**
	CA	**MLP**	0.3842	-0.0015	0.7700	0.0516
	CA	**MLR**	0.1013	-0.2845	0.4871	0.9758
	CA	**GP**	1.4582	1.0724	1.8440	**0.0000**
	LMT	**BN**	-0.9984	-1.3842	-0.6126	**0.0000**
	LMT	**MLP**	-0.1752	-0.5610	0.2105	0.7883
	LMT	**MLR**	-0.4582	-0.8440	-0.0724	**0.0093**
	LMT	**GP**	0.8987	0.5129	1.2845	**0.0000**
	BN	**MLP**	0.8232	0.4374	1.2089	**0.0000**
	BN	**MLR**	0.5402	0.1544	0.9260	**0.0009**
	BN	**GP**	1.8971	1.5113	2.2829	**0.0000**
	MLP	**MLR**	-0.2830	-0.6687	0.1028	0.2923
	MLP	**GP**	1.0740	0.6882	1.4597	**0.0000**
	MLR	**GP**	1.3569	0.9711	1.7427	**0.0000**
Girls	**CA**	**LMT**	-0.0796	-0.5303	0.3710	0.9961
	CA	**BN**	-0.2456	-0.6962	0.2051	0.6298
	CA	**MLP**	0.0708	-0.3799	0.5215	0.9977
	CA	**MLR**	0.0664	-0.3843	0.5170	0.9983
	CA	**GP**	0.7987	0.3480	1.2493	**0.0000**
	LMT	**BN**	-0.1659	-0.6166	0.2847	0.9011
	LMT	**MLP**	0.1504	-0.3002	0.6011	0.9330
	LMT	**MLR**	0.1460	-0.3046	0.5967	0.9407
	LMT	**GP**	0.8783	0.4277	1.3290	**0.0000**
	BN	**MLP**	0.3164	-0.1343	0.7670	0.3419
	BN	**MLR**	0.3119	-0.1387	0.7626	0.3582
	BN	**GP**	1.0442	0.5936	1.4949	**0.0000**
	MLP	**MLR**	-0.0044	-0.4551	0.4462	1.0000
	MLP	**GP**	0.7279	0.2772	1.1785	**0.0001**
	MLR	**GP**	0.7323	0.2816	1.1830	**0.0001**

The results obtained in this study showed that ML methods can be used to estimate bone age. In this study, a data set was created using numerical measurements related to the carpal area. In the next studies, it is planned to develop methods to estimate bone age and establish an automation system directly from radiological images. The development of an automation system that allows for predicting bone age over the image will both reduce the cost of processing and reduce dependence on the expert.

There are some limitations to our study. Due to the limitations of the database where time and samples were taken for each month group, approximately 10-15 girls and boys were examined. It is possible to increase the forecast performance by expanding the database. The proper location of the graphics examined in this study was an important factor. Since the dataset was formed in the bone areas in the carpal region, we had to exclude a large number of cases due to improper shooting techniques, since the wrong positioning in this region would adversely affect the success of the results.

Our study group was determined retrospectively. Patients diagnosed with developmental disorders were not included in the study. However, cases that are not clinically reported can be included in the study population. This may have affected the estimation performance of the methods, but it has been accepted that the effects did not make a significant difference between the methods.

In conclusion, the results obtained from this study showed that ML-based classification methods have high success in estimating bone age. While estimates based on expert opinion may differ according to the expert's experience and many other factors, ML-based approaches are more stable because they learn from data. Because of these properties, it is an important advantage that ML-based estimation of bone age can be adapted to any population. As a result, we propose that ML-based classification models can be used to estimate bone age. Furthermore, the results anticipate that the development of ML-based automation systems will reduce the reliance on expert opinion in bone age estimation.

Acknowledgments

This paper has been granted by the Muğla Sıtkı Koçman Üniversity Research Projects Coordination Office. Project Grant Number: 17/217 and title, “Esnek Hesaplama Yöntemleri İle Radyolojik Görüntülerden Kemik Yaşı Tahmini”.

Ethics Committee Approval: Ethics committee permission was obtained for this study with the protocol number 5514 of Muğla Sıtkı Koçman University Human Research Ethics Committee dated 13.03.2017.

References

1. Gilsanz V and Ratib O. Hand Bone Age: A Digital Atlas of Skeletal Maturity. 2005; 98. Springer Science & Business Media, Heidelberg.
2. Maggio A, Flavel A, Hart R, et al. Skeletal age estimation in a contemporary Western Australian population using the Tanner-Whitehouse method. Forensic Sci Int. 2016;63:1-8.
3. Pinchi V, De Luca F, Ricciardi F, et al. Skeletal age estimation for forensic purposes: A comparison of GP, TW2 and TW3 methods on an Italian sample. Forensic Sci Int. 2014;238:83-90.
4. Cantekin K, Çelikoğlu M, Miloglu O, et al. Bone Age Assessment: The Applicability of the Greulich-Pyle Method in Eastern Turkish Children. J Forensic Sci. 2012;57(3):679-82.
5. Öztürk F, Karataş OH, Mutaf IH, et al. Bone age assessment: comparison of children from two different regions with the Greulich–Pyle method In Turkey. Aust J Forensic Sci. 2016;48(6):694-703.
6. Büken B, Şafak AA, Yazıcı B, et al. Is the assessment of bone age by the Greulich–Pyle method reliable at forensic age estimation for Turkish children? Forensic Sci Int. 2007;173:146-53.
7. Berst MJ, Dolan L, Bogdanowicz MM, et al. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR Am J Roentgenol. 2001;176(2):507-10.
8. King DG, Steventon DM, O'sullivan MP, et al. Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods. Br J Radiol. 1994;67(801):848-51.
9. Guraksin GE, Uguz H, Baykan OK. Bone age determination in young children (newborn to 6 years old) using support vector machines. Turk J Elec Eng&Comp Sci. 2016;24:1693-708.
10. Gertych A, Zhang A, Sayre J, et al. Bone age assessment of children using a digital hand atlas. Comp Med Imaging Graph. 2007;31(4-5):322-31.
11. Pietka E, Pospiech-Kurkowskaa S, Gertych A, et al. Integration of computer assisted bone age assessment with clinical PACS. Comput Med Imaging Graph. 2003;27(2-3):217-28.
12. Seok J, Hyun B, Kasa-Vubu J, et al. Automated Classification System for Bone Age X-ray Images. IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE. 2012;208-13.
13. Tristan-Vega A, Arribas JI. A Radius and Ulna TW3 Bone Age Assessment System. IEEE Trans Biomed Eng. 2008;55(5):1463-76.
14. Liu J, Qi J, Liu Z, et al. Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method. Comput Med Imaging Graph. 2008;32(8):678-84.
15. Hasaltın E, Beşdok E. El-bi̇lek röntgen görüntüleri̇nden radyoloji̇k kemi̇k yaşı tespitinde yapay sinir ağları kullanımı. National Conference of Electrical, Electronics and Computer Engineering. 2004;8-12.
16. Thangam P, Mahendiran TV. Tetrolets-based System for Automatic Skeletal Bone Age Assessment. Int J Eng Res Sci. 2015;1:21-33.
17. Darmawan MF, Yusuf SM, Abdul Kadir MR, et al. Comparison on three classification techniques for sex estimation from the bone length of Asian children below 19 years old: An analysis using different group of ages. Forensic Sci Int. 2015;247:130.e1-11.
18. Lee JH, Kim KG. Applying Deep Learning in Medical Images:The Case of Bone Age Estimation. Healthc Inform Res. 2018;24(1):86-92.
19. Iglovikov I, Rakhlin A, Kalinin AA, et al. Paediatric bone age assessment using deep convolutional neural networks, Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. 2018; 300-8. Springer, Cham.
20. Hyunkwang L, Tajmir S, Lee J, et al. Fully Automated Deep Learning System for Bone Age Assessment. J Digit Imaging. 2017;30(4):427-41.
21. Spampinatoa C, Palazzoa C, Giordano D, et al. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal. 2017;36:41-51.
22. Thodberg HH, Kreiborg S, Juul A, et al. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging. 2009;28 (1):52-66.
23. Predicting Skeletal Age avaible at: https://www.16bit.ai/bone-age
24. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.
25. Rumelhart DE, Geoffrey EH, Ronald JW. Learning internal representations by error propagation. No. ICS-8506. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
26. Quinlan JR. Simplifying decision trees. Int J Man Mach Stud. 1987;27:221-34.
27. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn. 1997;29:131-63.
28. Hosmer DW, Stanley JL, Sturdivant RX. Applied logistic regression. 2013;398. John Wiley & Sons.
29. Weka 3: Data Mining Software in Java avaible at: https://www.cs.waikato.ac.nz/ml/weka/
30. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32:675-701.
31. Friedman M. A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat. 1940;11:86-92.
32. Korting TS. C4. 5 algorithm and multivariate decision trees. Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil 2006.
33. Friedman JH, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat. 2000;28:337-407.
34. Godbole S, Sarawagi S, Chakrabarti S. Scaling multi-class support vector machines using inter-class confusion. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002.
35. Koc A, Karaoglanoglu M, Erdogan M, et al. Assessment of bone ages: is the Greulich-Pyle method sufficient for Turkish boys? Pediatr Int. 2001;43(6):662-5.