Original Article / Özgün Araştırma

Meandros Med Dent J
doi:10.4274/meandros.galenos.2023.22043
Meandros Med Dent J 2023;24(2):161-167

161

Comparison of Machine Learning Methods to Predict Incomplete Atypical Femoral Fracture After Bisphosphonate Use in Postmenopausal Women

Menopoz Sonrası Kadınlarda Bifosfonat Kullanımı Sonrası Tamamlanmamış Atipik Femur Kırıklarının Tahmini için Makine Öğrenmesi Modellerinin Karşılaştırılması

Sultan Turhan[1], Tuğba Dübektaş Canbek[2], Umut Canbek[3], Eralp Doğu[1]

[1]Muğla Sıtkı Koçman University Faculty of Sciences, Department of Statistics, Muğla, Turkey
[2]Mugla Sıtkı Koçman University Faculty of Medicine, Department of Internal Medicine, Muğla, Turkey
[3]Muğla Sıtkı Koçman University Faculty of Medicine, Department of Orthopedics and Traumatology, Muğla, Turkey

Abstract

Objective: Long-term use of bisphosphonates (BP) for treating osteoporosis may cause incomplete atypical femoral fracture. In this study, we compared the classification and risk estimation of incomplete atypical femoral fractures, which is an alternative approach to clinical risk assessment.

Materials and Methods: A data set was randomly selected from women using postmenopausal BP. We identified a class imbalance problem in the population and created a balanced structure using the density-based synthetic minority over-sampling technique. We compared machine learning algorithms and conducted a case study.

Results: We solved the class imbalance problem with the density-based synthetic minority over-sampling technique and found that the random forest and adaboost methods achieved the highest performance in the classification step.

Conclusion: It is recommended to apply resampling methods in cases where there is an unbalanced class problem such as incomplete atypical femoral fracture. Ensemble methods perform better than traditional methods in this study.

Keywords: Classification, unbalanced data, disease diagnosis, orthopedics, incomplete atypical fractures

Öz

Amaç: Bifosfonatların (BP) osteoporoz tedavisinde uzun süreli kullanımı tam olmayan atipik femur kırığına neden olabilir. Bu çalışmada, tamamlanmamış atipik femur kırıklarının sınıflandırılması ve risk tahmini için gelişmiş makine öğrenimi modellerinin performansını karşılaştırmayı amaçlanmaktadır.

Gereç ve Yöntemler: Veri seti, menopoz sonrası BP kullanan kadınların rastgele bir alt kümesini içerir. Popülasyonda bir sınıf dengesizliği sorunu belirledik ve yoğunluğa dayalı sentetik azınlık aşırı örnekleme tekniği kullanarak dengeli bir yapı oluşturduk. Makine öğrenmesi algoritmalarını karşılaştırdık ve bir olgu çalışması gerçekleştirdik.

Bulgular: Bu çalışmada, geleneksel lojistik regresyon yaklaşımını birkaç gelişmiş topluluk öğrenme yöntemiyle karşılaştırılmıştır ve rastgele orman ve Adaboost yöntemlerinin en iyi tahmin performansını elde ettiği sonucuna varılmıştır.

Sonuç: Bu çalışmada, atipik femur kırığını tahmin etmek için tekrarlanabilir bir makine öğrenimi iş akışı gösterilmiştir. Gelişmiş tahmine dayalı modeller, geleneksel modellerle karşılaştırılmış ve bunların geleneksel modellerden daha iyi performans gösterdiğini gösterilmiştir.

Anahtar Kelimeler: Sınıflandırma, dengesiz veri, hastalık tanısı, ortopedi, tamamlanmamış atipik kırıklar

Address for Correspondence/Yazışma Adresi: Eralp Doğu, Muğla Sıtkı Koçman University Faculty of Sciences, Department of Statistics, Muğla, Turkey
E-mail: eralp.dogu@mu.edu.tr
ORCID ID: orcid.org/0000-0002-8256-7304

Received/Geliş Tarihi: 11.02.2022
Accepted/Kabul Tarihi: 11.06.2023

Introduction

Osteoporosis is an important orthopedic disease that occurs with low bone density and deterioration of bone structure (1-3). Early diagnosis of osteoporosis and appropriate and effective treatment such as bisphosphonates (BP) is very important to prevent potential fractures (4). However, long-term use of BP in the treatment of osteoporosis can result in incomplete atypical femoral fractur (iAFF) (5). These fractures are life-threatening (6). Therefore, determining the risk of fracture can be a great difficulty (7). More recently, machine learning (ML) methods have been increasingly utilized as they provide robust and versatile means of risk prediction for various medical domains. Although ML is relatively new to the field of orthopedics, it is essential for researchers in this field to fully understand ML (8,9).

In this study, we present a comparison of ML algorithms to risk predict iAFFs in the post-menstrual period. In our previous study, we had some difficulties in identifying risk factors due to numerical imbalances between groups (10). From this point of view, it is aimed to develop ML in data with class imbalance between groups in investigating risk factors in diseases with low prevalence. Using existing ML algorithms without considering data preprocessing to balance data sets makes it very difficult to develop an effective model. To prevent this, oversampling methods are applied before the training step in the application. Here, we first applied DBSMOTE to the training dataset with class imbalance problem because it is easy to extract information from small datasets and is needed in the real world. Next, we made a comparison between traditional learning methods and advanced ML methods in terms of evaluation criteria such as accuracy, sensitivity, specificity and kappa. To test the interpretability of the model that was selected after comparisons, we conducted a validation study for two patients.

Materials and Methods

Data

Study Population

This study was carried out in Menteşe district of Muğla province, where a population. The national health registry showed that 2,746 postmenopausal women in this region were using BP. Inclusion criteria were defined as age >50, female gender, diagnosis of osteoporosis and duration of BP use (BITIME). The true prevalence of iAFF is unknown, as some patients with iAFF are asymptomatic and do not seek treatment. We assumed 10% prevalence of iAFF and used 5% α significance and ±5% precision levels. As a result, the study continued with 132 patients and iAFF was detected in 14 of these 132 patients. Imbalanced ratio was 8.42 showing a moderate imbalance.

The data were obtained during the research project 17/064 supported by the Scientific Research Projects unit of Muğla Sıtkı Koçman University and ethics committee approval was obtained between 08.2016-08.2017 (approval number: 2016/55, date: 17.06.2016). Informed consent was obtained from all participants included in this study. Table 1 shows the descriptive statistics of the iAFF data set.

Imaging Studies

In order not to miss the early insufficiency fracture, whole body bone scintigraphy and anteroposterior and lateral radiographs were taken from all subjects. The presence of increased involvement in the lateral cortex of the femur by bone scintigraphy, focal changes in the lateral cortex of the femur (radiolucent line, focal-generalized cortical thickening, lines, cavities) were accepted as an iAFF. A consensus diagnosis of an iAFF was made with both radiographic and scintigraphic images (a team consisting of 2 orthopedists, 1 radiology and 1 nuclear medicine specialist). Although the ASBMR case definition does not include a bone scan or MRI, many authors suggest that an advanced imaging modality can be used for definitive diagnosis if there is a high suspicion of an iAFF (6).

Laboratory Tests

All measurements of the patients are given in Table 1. Bone mineral density (BMD) measurements of the femoral neck and anteroposterior lumbar spine were made using a dual-energy X-ray absorptiometry machine. According to World Health Organization criteria, normal BMD was defined as less than 1 standard deviation (SD) of young adult peak BMD (T-score), osteopenia was defined as a value between 1.0 and 2.5 SD. Young adult peak BMD and osteoporosis were defined as a value greater than or equal to 2.5 SD of young adult peak BMD.

Statistical Analysis

ML is defined as a multidisciplinary area that uses statistics, mathematics and computer science. It focuses on building models that learn from data and increase accuracy over time. ML can be thought of as deciding which treatment is most effective for a patient with certain characteristics, providing more accurate answers to clinical questions such as the expected risks/benefits in the short and long term for a specific disease (11,12). They can be performed using approaches such as traditional and advanced learning (13,14). In this study, we used logistic regression, decision tree, random forest, adaptive boosting, extreme gradient boosting methods

Logistic Regression (LR): LR is a traditional statistical learning method that uses a logistic link function to model a bilateral orthopedic outcome based on patient-level risk factors (15).

Decision Tree (DT): DT is a tree-based algorithm consisting of a series of decision tests that work with the divide and conquer method. Thanks to the tree structure, it makes it easier for experts to interpret the model and to detect high-risk patients (16).

Random Forest (RF): RF is an ensemble method (bagging) created from decision trees. By combining more than one decision trees in the RF, a decision forest is created and the final estimation of the patient risk is made by combining the estimation results obtained from each decision tree (17).

Adaptive Boosting (ADABOOST): ADABOOST is an ensemble method (boosting) proposed by Freund and Schapire (18). ADABOOST initially starts with an even distribution for each sample and finds the weakest classifier based on classification performance. Then, it updates the weights, focusing on weakly classified samples. By combining weak classifiers as a result of a certain iteration, a strong classifier is created for disease classification (17). **Extreme Gradient Boosting (XGBOOST):** Gradient boosting is an ensemble method (boosting) that creates a prediction model for classification problems. XGBOOST builds and generalizes the model iteratively as other incremental methods. One of the most important features that distinguishes this method from others is its extra randomization parameter can be used to reduce the correlation between trees (19).

Density-Based Synthetic Minority Over-sampling Technique (DBSMOTE): DBSMOTE, Bunkhumpornpat et al. (20). It is based on the oversample randomly shaped set developed

Table 1. Descriptive statistics for iAFF dataset raw and DBSMOTE

		RAW data	DBSMOTE
Feature abbreviations	**Feature descriptions**	**Mean ± SD/n**	**Mean ± SD/n**
Characteristic			
Age (years)	Patient's age	72.79±7.35	73.67±7.76
Height (cm)	Patient height	149.02±5.23	149.70±0.04
Weight (kg)	Patient's weight	63.82±12.11	64.91±10.25
BMI (kg/m^2)	Body mass index	28.71±5.14	28.93±4.21
Medication			
BITIME	Bisphosphonate usage time	7.71±3.40	8.74±3.92
Steroid	Steorid usage history		
Present		19	13
Absent		113	85
PPI	Proton pump inhibitor		
Present		57	42
Absent		75	56
DM	Diabetes mellitus		
Present		25	18
Absent		107	81
Thyroid	Thyroid status		
Normal		109	74
Hypothyroidism		12	14
Hyperthyroidism		11	8
Test result			
DVIT (ng/mL)	Vitamin D level	30.24±12.81	30.21±12.89
PTH (pg/mL)	Parathyroid hormone level	58.12±27.39	68.18±28.78
ALP (U/L)	Alkaline phosphatase	65.05±20.38	66.11±26.58
HIPTS	Hip T-score	-1.78±0.77	-1.80±0.68
Vertebrats	Vertebra T-score	-2.49±1.13	-2.48±0.89

BMI: Body mass index, SD: Standard deviation, DM: Diabetes mellitus, PTH: Parathyroid hormone, ALP: Alkaline phosphatase

by DBSCAN. The purpose of the DBSMOTE algorithm is to try to solve the class imbalance problem by adaptively generating synthetic new examples from the minority class through linear interpolation between existing minority class instances. DBSMOTE aims to reduce the bias in SMOTE and can adaptively change the decision boundary to focus on hard-to-learn samples (21,22).

Evaluation Criteria Methods Advocated in the Paper

Sensitivity ($TP/((TP + FN))$) is the ratio of predicted positive class values (TP) to all positive class values (TP+FN) (23,24). Specificity ($TP/((TP + FN))$) is the ratio of correctly predicted negative class values (TN) to all negative class values (23,24). Precision [$TP/((TP+FP))$] is the ratio of correctly predicted positive class value (TP) to all positively predicted class values (23,24). Balanced accuracy and kappa values were also considered.

Proposed ML Workflow

The steps of the proposed ML workflow are given in Figure 1.

The applications were conducted in R, an open-source software. During the study, caret, caretEnsemble, smotefamily, ggplot2, gridExtra and lime packages were used. Data set is divided into two parts as 25% test and 75% training. In this study DBSMOTE was also applied. Here, we install a 5-fold cross-validation approach to avoid overfitting. In the classification with raw and post-DBSMOTE data set, random search tunelength =5 was made to adjust the hyper parameters of the models. To increase the interpretability of the mode we use the lime package for interpretation purposes.

Results

Class Imbalance Reduces the Performance of ML Methods to Effectively Detect iAFFs after BP Use

The class imbalance problem directly affects the performance criteria in the application of ML methods. Table 2 shows the results per model in the presence of class imbalance.

LR (balanced accuracy =0.45, kappa =-0.07, sensitivity =0, specificity =0.90, F1=NA, Precision =0) was observed to have the lowest performance results among the other methods. XGBOOST (balanced accuracy =0.90, kappa =35, sensitivity =0.50, specificity =0.98, F1=40, Precision=0.33) compared to other methods although it is good in balanced accuracy and kappa, it is still not sufficient for other criteria.

DBSMOTE Helps ML Methods to Effectively Detect iAFFs after BP Use

In this step, we applied DBSMOTE and observed a significant improvement in the performance of models shown in Table 2.

RF (balanced accuracy =0.88, kappa =0.79, sensitivity =1.00, specificity =0.77, F1=0.91, precision =0.84) method was found to provide higher accuracy than other methods. Classification success order after RF is followed by LR (balanced accuracy =0.73, kappa =0.47, sensitivity =0.88, specificity =0.58, F1=0.77, Precision =0.69), ADABOOST (balanced accuracy =0.88, kappa =0.78, sensitivity =0.96, specificity =0.80, F1=0.92, Precision=0.87), DT (balanced accuracy =0.80, kappa =0.65, sensitivity =0.90, specificity =0.70, F1=0.85, precision =0.81) and XGBOOST (balanced accuracy =0.84, kappa =0.70, sensitivity =0.93, specificity =0.75, F1=0.88, precision =0.84) methods.

The very low classification success of the diagnosis of iAFF made with raw data was interpreted as the result of imbalance. When these results are compared, it shows that LR method is relatively inadequate compared to RF, ADABOOST and XGBOOST methods in the presence of class imbalance. Another point to note here is that the XGBOOST method performs significantly higher than other methods in the presence of class imbalance. When the results after DBSMOTE are examined, it is observed that the RF and ADABOOST methods over performs others.

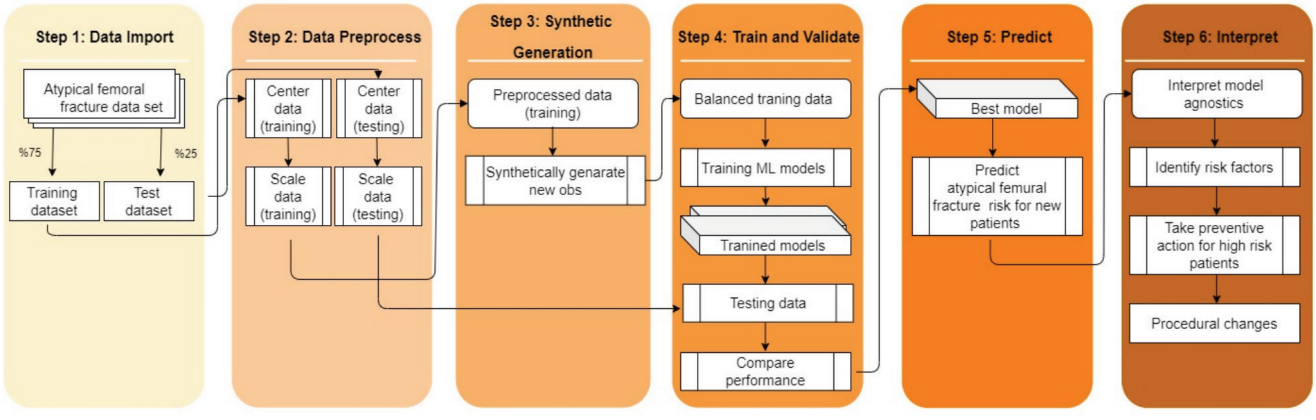

Figure 1. Proposed ML workflow for risk prediction of femoral fracture

ML: Machine learning

ML Methods Help Effectively Interpret the Risk Factors for iAFFs after BP Use

Fracture risk assessment can guide clinicians and individuals in understanding the risk of having a fracture and speed up the decision-making process to reduce these risks. Figure 2 shows the order of the most important risk factors based on RF method.

As Figure 3 we also conducted a study of RF risk assessment on two test cases where the labels were not provided. In this case study, the characteristics that contribute to the classification of each patient and the decision rules for these characteristics were determined. After the RF model was trained, randomly selected patients #7 and #119 were classified for iAFFs. Here the y-axis gives the decision rules, and the X-axis gives the weights in this decision (Figure 3).

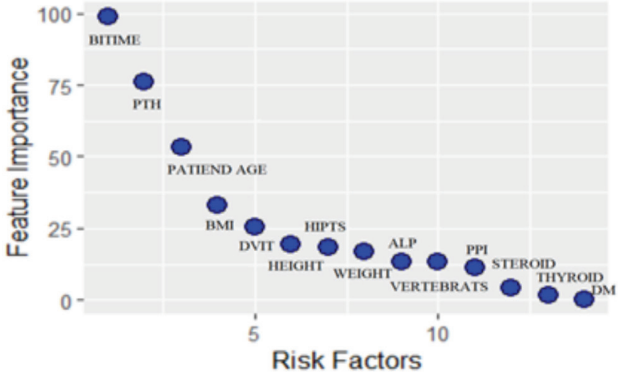

Figure 2. Feature importances based on RF

RF: Random forest

Here, patient 7 was assigned class 0-label with probability 0.96 and patient 1 was assigned 1 labeled class with probability 0.86. When the result of the 7th patient is examined, the fact that BITIME <=6, PTH <=41.5 and weight <=54.6 contribute to the patient entering the no-risk class. 10.2< BITIME and HIPTS <=-2.3 contributed to the classification of patient #119 as involvement, while BMI <=24.6 does not support the presence of involvement.

The findings obtained after this case study were approved by the orthopedic and traumatology specialist, and it was concluded that the 7th and 119th patients were in the classes we predicted.

Conclusion

Today, an orthopedist needs accurate predictions of the outcome of their patients' disease, and therefore high-performance methods are vital to support treatment decisions. The percentage of people in the geatric age group in the general population is increasing and the use of BP group drugs for the prevention of osteoporotic fractures in this age group is becoming widespread (4). In the investigation of risk factors for iAFF in patients using BP, numerical group imbalances emerge between iAFF and non-iAFF groups. It is a common problem in studies to determine the risk factors associated with such diseases with low prevalence values. Early detection of iAFF significantly changes patient mortality and morbidity. In this respect, it is important to reveal the risk factors and true prevalence. As an important decision support tool, machine learning methods are used to potentially transform large medical data sets int

Table 2. Performance result after classification with the raw data set and performance results after classification with DBSMOTE. Darker color codes show better performance

Metric (raw data)	LR	RF	DT	ADABOOST	XGBOOST
Balanced accuracy	0.45	0.50	0.48	0.48	0.90
Kappa	-0.07	-0.04	-0.04	-0.04	0.35
Sensitivity	0	0	0	0	0.50
Specificity	0.90	0.90	0.96	0.96	0.98
F1	NA	NA	NA	NA	0.40
Precision	0	0	0	0	0.33
Metric (DBSMOTE)					
Balanced accuracy	0.73	0.88	0.80	0.88	0.84
Kappa	0.47	0.79	0.65	0.78	0.70
Sensitivity	0.88	1.00	0.90	0.96	0.93
Specificity	0.58	0.77	0.70	0.80	0.75
F1	0.77	0.91	0.85	0.92	0.88
Precision	0.69	0.84	0.81	0.87	0.84

LR: Logistic regression, RF: Random forest, DT: Decision tree

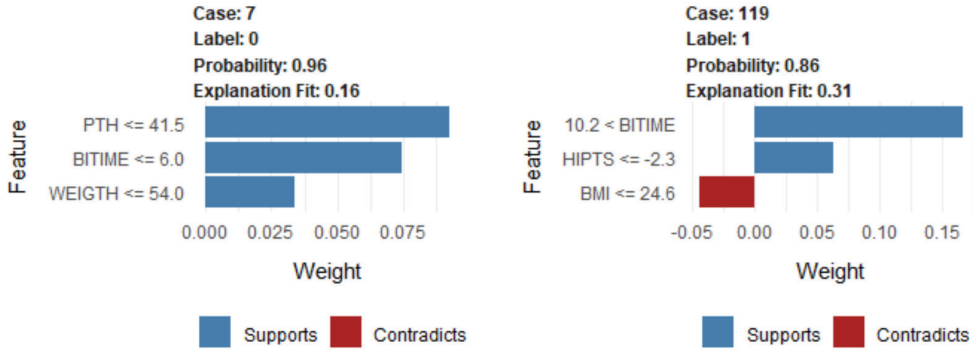

Figure 3. Model agnostics of the classification procedure for the 7th and 119th patient in the iAFF test data set

Into a meaningful and efficient structure (9,13-14,25). However, ML methods are affected by the class imbalance problem and this problem has a big impact on classification performance. LR, DT, RF, ADABOOST and XGBOOST methods were used in this study. We observed that the LR method was highly affected by the class imbalance problem and that the DT, RF, ADABOOST and XGBOOST methods also achieved low classification success. Therefore, the data set is balanced with DBSMOTE to eliminate the class imbalance problem. The balanced data set was integrated into LR, DT, RF, ADABOOST and XGBOOST methods. The results showed that the RF and ADABOOST methods was the best performing method among these algorithms. Then, risk factors were determined by conducting a case study using RF method. When evaluating the risk of iAFF, we considered risk factors. We have reported varying degrees of severity of the RF method to help better understand key risk factors. Besides, we found that risk factors such as the duration of BP use, PTH level, age, DVIT level and body mass index contribute significantly to the high fracture risk.

Our study has some limitations. At the data level, the study was conducted in a small number of cohorts, and the data set used was obtained from patients in a particular region. In addition, this study provides cost effective determination of the risk factors of the disease without the need for X-ray and scintigraphy in the field of orthopedics.

Ethics

Ethics Committee Approval: The data were obtained during the research project 17/064 supported by the Scientific Research Projects unit of Muğla Sıtkı Koçman University and ethics committee approval was obtained between 08.2016-08.2017.

Informed Consent: Informed consent was obtained from all participants included in this study.

Peer-review: Externally peer-reviewed.

Authorship Contributions

Concept: S.T., U.C., E.D., Design: S.T., U.C., E.D., Data Collection or Processing: T.D.C., U.C., E.D., Analysis or Interpretation: S.T., T.D.C., U.C., E.D., Literature Search: S.T., U.C., E.D., Writing: S.T., T.D.C., U.C., E.D.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The data were obtained during the research project 17/064 supported by the Scientific Research Projects unit of Muğla Sıtkı Koçman University.

References

1. Fischer S, Kapinos KA, Mulcahy A, Pinto L, Hayden O, Barron R. Estimating the long-term functional burden of osteoporosis-related fractures. Osteoporos Int 2017; 28: 2843-51.
2. Canbek U, Hazer DB, Rosberg H, Akgün U, Canbek TD, Cömert A, et al. The Effect of Bisphosphonates on Lumbar Vertebral Disc Height. Ege Klinikleri Tıp Dergisi 2019; 57: 52-6.
3. Hopkins RB, Tarride JE, Leslie WD, Metge C, Lix LM, Morin S, Finlayson G, et al. Estimating the excess costs for patients with incident fractures, prevalent fractures, and nonfracture osteoporosis. Osteoporos Int 2013; 24: 581-93.
4. Silverman S, Christiansen C. Individualizing osteoporosis therapy. Osteoporos Int 2012; 23: 797-809.
5. Nishino T, Hyodo K, Matsumoto Y, Yanagisawa Y, Yoshizawa T, Yamazaki M. Surgical results of atypical femoral fractures in long-term bisphosphonate and glucocorticoid users - Relationship between fracture reduction and bone union. J. Orthop 2020; 19: 143-9.
6. Shane E, Burr D, Abrahamsen B, Adler RA, Brown TD, Cheung AM, et al. Atypical subtrochanteric and diaphyseal femoral fractures: second report of a task force of the American Society for Bone and Mineral Research. J Bone Miner Res 2014; 29: 1-23.
7. WSG on the P. and M. of Osteoporosis, Prevention and management of osteoporosis : report of a WHO scientific group. World Health Organization. Geneva PP, 2003 (Online). Available: https://apps.who.int/iris/handle/10665/42841.
8. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine 2016; 375: 1216-9.
9. Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. 2018; 6.
10. Canbek U, Akgun U, Soylemez D, Canbek TD, Aydogan NH. Incomplete atypical femoral fractures after bisphosphonate use in postmenopausal women. J Orthop Surg 2019; 27: 1-10.

11. Mitchell TM, Machine Learning. 1st. New York: McGraw-Hill; 1997:414.
12. Öztürk H, Türe M, Kıylıoğlu N, Kurt Ömürlü İ. The Comparison of Different Dimension Reduction and Classification Methods in Electroencephalogram Signals. Meandros Med Dent J 2018; 19: 336-44.
13. Kruse C, Eiken P, Vestergaard P. Machine Learning Principles Can Improve Hip Fracture Prediction. Calcif Tissue Int 2017; 100: 348-60.
14. Engels A, Reber KC, Lindlbauer I, Rapp K, Büchele G, Klenk J, et al. Osteoporotic hip fracture prediction from risk factors available in administrative claims data-A machine learning approach. PLoS One 2020; 15: 1-14.
15. Berkson J. Application of the logistic function to bio-assay. J Am Stat Assoc 1944; 39: 357-65.
16. Tu PL, Chung JY. A new decision-tree classification algorithm for machine learning. Proc - Int Conf Tools with Artif Intell ICTAI 1992; 370-7.
17. Breiman L. Random Forests. Mach Learn 2001; 45: 5-32.
18. Freund Y, Schapire RRE. Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference 1996; 148-56.
19. Rusdah DA, Murfi H. XGBoost in handling missing values for life insurance risk prediction. SN Appl Sci 2020; 8: 1-10.
20. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. DBSMOTE: Density-based synthetic minority over-sampling technique. Appl Intell 2012; 36: 664-84.
21. Pun S, Thapa S, Timilsina S, Customer Churn Prediction Using ADASYN Sampling Technique and Ensemble Model. Proc IOE Grad Conf 2019; 6: 513-8.
22. Wang JB, Zou CA, Fu GH. AWSMOTE: An SVM-Based Adaptive Weighted SMOTE for Class-Imbalance Learning, Sci. Program 2021; 9947621:1-9947621:18.
23. Turhan S, Tunç M, Doğu E, Balcı Y. Machine learning in forensic science and forensic medicine: Research on the literature. Adli Tıp Dergisi 2022; 36: 1-7.
24. Wu H, Meng FJ. Review on evaluation criteria of machine learning based on big data. In Journal of Physics: Conference Series 2020; 1486: 5; 052026.
25. Turhan S, Özkan Y, Yürekli BS, Suner A, Doğu E. Comparison of Ensemble Learning Methods for Disease Diagnosis in Presence of Class Unbalanced: Case of Diabetes. Turkiye Klin J Biostat 2020; 12: 16-26.