ORIGINAL ARTICLE

Detection and classification of femoral neck fractures from plain pelvic X-rays using deep learning and machine learning methods

Hüseyin Fatih Sevinç,[1] Kemal Üreten,[2] Talha Karadeniz,[2] Gökhan Koray Gültekin[3]

[1]Department of Orthopedics and Traumatology, Versa Hospital, Kapadokya University, Nevşehir-*Türkiye*

[2]Department of Computer Engineering, Çankaya University, Ankara-*Türkiye*

[3]Department of Electrical and Electronics Engineering, Yıldırım Beyazıt University, Ankara-*Türkiye*

ABSTRACT

BACKGROUND: Femoral neck fractures are a serious health concern, particularly among the elderly. The aim of this study is to diagnose and classify femoral neck fractures from plain pelvic X-rays using deep learning and machine learning algorithms, and to compare the performance of these methods.

METHODS: The study was conducted on a total of 598 plain pelvic X-ray images, including 296 patients with femoral neck fractures and 302 individuals without femoral neck fractures. Initially, transfer learning was applied using pre-trained deep learning models: VGG-16, ResNet-50, and MobileNetv2.

RESULTS: The pre-trained VGG-16 network demonstrated slightly better performance than ResNet-50 and MobileNetV2 for detecting and classifying femoral neck fractures. Using the VGG-16 model, the following results were obtained: 95.6% accuracy, 95.5% sensitivity, 93.3% specificity, 95.7% precision, 95.5% F1 Score, a Cohen's kappa of 0.91, and the Receiver Operating Characteristic (ROC) curve of 0.99. Subsequently, features extracted from the convolution layers of VGG-16 were classified using common machine learning algorithms. Among these, the k-nearest neighbor (k-NN) algorithm outperformed the others and exceeded the accuracy of the VGG-16 model by 1%.

CONCLUSION: Successful results were obtained using deep learning and machine learning methods for the detection and classification of femoral neck fractures. The model can be further improved through multi-center studies. The proposed model may be especially useful for physicians working in emergency departments and for those not having sufficient experience in evaluating plain pelvic radiographs.

Keywords: Deep learning; feature extraction; femoral neck fractures; machine learning; pre-trained networks.

INTRODUCTION

Hip fractures are becoming increasingly common as the average age of the population rises, and they represent a serious health problem, particularly among the elderly. Early diagnosis is critical in cases of hip fracture, as delays in diagnosis lead to complications such as malunion, osteonecrosis, and arthritis, and can increase morbidity and mortality rates.[1-3] The one-month mortality rate for hip fractures is around 10%, and the risk of complications, including infections and thromboembolic events, increases when surgical treatment is delayed.[4-6]

Plain pelvic X-rays are a simple, cost-effective, easily accessible, and rapid imaging method for diagnosing femoral neck fractures. However, studies report that approximately 2% of

Cite this article as: Sevinç HF, Üreten K, Karadeniz T, Gültekin GK. Detection and classification of femoral neck fractures from plain pelvic X-rays using deep learning and machine learning methods. Ulus Travma Acil Cerrahi Derg 2025;31:783-788.

Address for correspondence: Hüseyin Fatih Sevinç
Department of Orthopedics and Traumatology, Versa Hospital, Kapadokya University, Nevşehir, Türkiye
E-mail: drhfatihsevinc@gmail.com

Ulus Travma Acil Cerrahi Derg 2025;31(8):783-788 DOI: 10.14744/tjtes.2025.75806
Submitted: 07.02.2025 Revised: 14.02.2025 Accepted: 15.06.2025 Published: 05.08.2025

all hip fractures cannot be identified on plain pelvic X-rays, and the misdiagnosis rate can be as high as 14%. While magnetic resonance imaging (MRI) and computed tomography (CT) offer higher sensitivity in detecting femoral neck fractures, these imaging methods are costly, may not be suitable for all patients, and are not available in every medical center.[7,8]

Convolutional neural networks (CNN), a deep learning method, consist of varying numbers of convolution layers, pooling layers, and one or two fully connected layers. CNNs are highly effective in image processing, feature extraction, and image classification, and they require little to no preprocessing. In the convolution layer, feature maps are generated by applying different filters to the input image. Depending on the number of convolution layers used, basic features, simple models, and advanced models can be derived from these feature maps. The goal is to achieve effective learning by progressing from low-level to high-level features. The pooling layer reduces the size of the feature maps produced by the convolution layer, and final classification is performed in the fully connected layer.[9]

The performance of CNNs can be enhanced through transfer learning and data augmentation methods. Transfer learning involves using pre-trained models, such as AlexNet, VGG-16, ResNet-50, and MobileNet, trained on the ImageNet dataset, for a new task.[10] In data augmentation, geometric and photometric transformations are applied to the training data. Geometric transformations include flipping, cropping, scaling, and rotating, while photometric transformations involve changes in lighting, color, and color jittering. Studies have shown that geometric transformations tend to be more effective than photometric ones.[11,12]

In recent years, many successful studies have been conducted using CNNs on various medical images. Several studies have employed CNNs and transfer learning methods for diagnosing hand fractures, foot fractures, and rib fractures from plain radiographs.[10,13,14] Additionally, some studies have applied deep learning and transfer learning techniques for the diagnosis of sacroiliitis, spina bifida occulta, and femoroacetabular impingement using plain pelvic X-rays.[15-17] Similarly, recent studies have utilized deep learning methods for the diagnosis and detection of hip fractures using plain pelvic X-rays.[18-20]

The aim of this study is to develop a method that will assist physicians in emergency departments and those with limited experience in evaluating plain hip X-rays in detecting femoral neck fractures, thereby reducing the reliance on advanced imaging modalities such as MRI and CT. To achieve this, transfer learning will first be applied using pre-trained CNN models. Then, feature extraction will be performed using these deep learning models, and the extracted features will be classified using machine learning algorithms. Finally, the performances of deep learning and machine learning methods will be compared.

MATERIALS AND METHODS

Dataset

Ethics committee approval for the study was obtained from the TC Nevşehir Hacı Bektaş Veli University Ethics Committee (Date: 19.04.2023, No: 2023/04). All procedures performed in this study involving human participants were conducted in accordance with the ethical standards of the institutional and/or national research committees, and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. This is a retrospective study conducted using patient pelvic radiographs. As no names or identifying information were used, informed consent was not obtained; this was stated in the ethics committee application and approved accordingly. The study was conducted using plain pelvic radiographs of patients who were admitted to the Emergency Department and Orthopedics and Traumatology Clinics of Nevşehir State Hospital between January 2017 and December 2022 due to hip trauma and were treated with a diagnosis of femoral neck fracture. All radiographs were reclassified by an Orthopedics and Traumatology specialist with more than five years of experience. A total of 598 plain pelvic X-rays were included in the study, comprising 296 patients with femoral neck fractures and 302 individuals without femoral neck fractures.

Data Processing Environment

In this study, the Python 3.9 programming language was used within a TensorFlow and Keras environment. Initially, the transfer learning method was applied using pre-trained networks: VGG-16, ResNet-50, and MobileNetV2. The hyperparameters used in this study were: optimizer — Adam; mini-batch size — 16; and number of epochs — 25. Feature extraction was then performed from the intermediate layers of the pre-trained VGG-16 model. These extracted features were classified using machine learning algorithms. Hyperparameter optimization for the machine learning algorithms was performed using the grid search method.

Data Preprocessing and Splitting

The plain pelvic X-rays used in the study varied in dimensions, and some included artifacts such as the patient's name, directional markers, or clothing. The radiographs were cropped to include the acetabulum, femoral head, femoral neck, trochanters, and the proximal one-fourth of the femur. All images were then resized to 224 × 224. Figure 1 shows examples of the original plain pelvic X-rays, while Figure 2 presents examples of the cropped and resized versions. For model evaluation, 15% of the dataset was reserved for testing. This test data was not used during the training or validation phases. Data augmentation was performed by applying a 30-degree right-left rotation, horizontal flip, and width and height shift ranges of 0.1 to the training radiographs. Table 1 shows the number of X-rays used for training, validation, and testing in this study.

Table 1. Numbers of pelvic X-rays used for training, validation, and testing

	Training	Validation	Testing	Total
Fracture	214	37	45	296
No Fracture	218	38	46	302
Total	432	75	91	

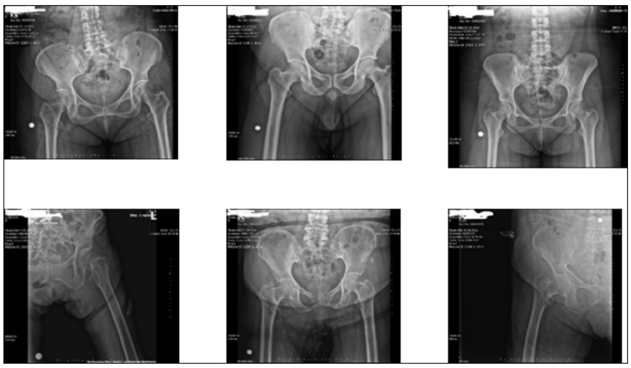

Figure 1. Original pelvic X-ray samples (top: unfractured; bottom: fractured).

Feature Extraction

Feature extraction is a necessary step for image classification using both machine learning and deep learning algorithms. It involves transforming visual elements such as edges, corners, textures, and colors into vector data. While traditional machine learning methods rely on hand-crafted feature extraction techniques, CNNs can serve as both classifiers and feature extractors in image processing. In CNNs, feature extraction is performed automatically. CNNs are highly effective at this task, with features extracted through convolutional layers or pooling layers added to them. The number of extracted features varies depending on the convolutional layers used. In this study, feature extraction was performed from an intermediate layer of the VGG-16 model, resulting in 512 extracted features.[21]

Statistical Analysis

In classification problems, a confusion matrix is used to compare predicted values with actual values. From this matrix, metrics such as accuracy, sensitivity, specificity, precision, F1 score, and Cohen's kappa are calculated using the values of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The area under the curve (AUC) is derived from the receiver operating characteristic (ROC) curve obtained during model testing. These performance metrics were calculated using the Python scikit-learn library. The following formulas were used:

Figure 2. Cropped and resized pelvic X-ray samples (top: unfractured; bottom: fractured).

RESULTS

In this study, conducted using plain pelvic X-rays for the detection of femoral neck fractures, the transfer learning method was applied using pre-trained VGG-16, ResNet-50, and MobileNetV2 networks. The VGG-16 model achieved an accuracy of 95.6%, with a 95% confidence interval (CI) ranging from 0.94 to 0.96 (alpha=0.95, CI: 0.94-0.96). The model also achieved 95.5% sensitivity (alpha=0.95, CI: 0.94-0.96), 93.3% specificity (alpha=0.95, CI: 0.91-0.94), 95.7% precision (alpha=0.95, CI: 0.94-0.97), 95.5% F1 score (alpha=0.95, CI: 0.94-0.96), a Cohen's kappa of 0.91, and an AUC of 0.99. Table 2 presents the performance metrics of all three deep learning models on the test data. Figure 3 displays the confusion matrix and ROC curve generated using the VGG-16 model.

The intermediate layer of the pre-trained VGG-16 network was used for feature extraction, from which 512 features were obtained. These extracted features were classified us-

Table 2. Performance metrics of pre-trained VGG-16, ResNet-50, and MobileNetV2 networks in femoral neck fracture classification using test data

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	Cohen's Kappa	AUC
VGG-16	95.6	95.5	93.3	95.7	95.5	0.91	0.99
ResNet-50	93.4	93.4	97.7	93.7	93.3	0.86	0.97
MobileNetV2	93.4	93.3	91.1	93.5	93.3	0.86	0.94

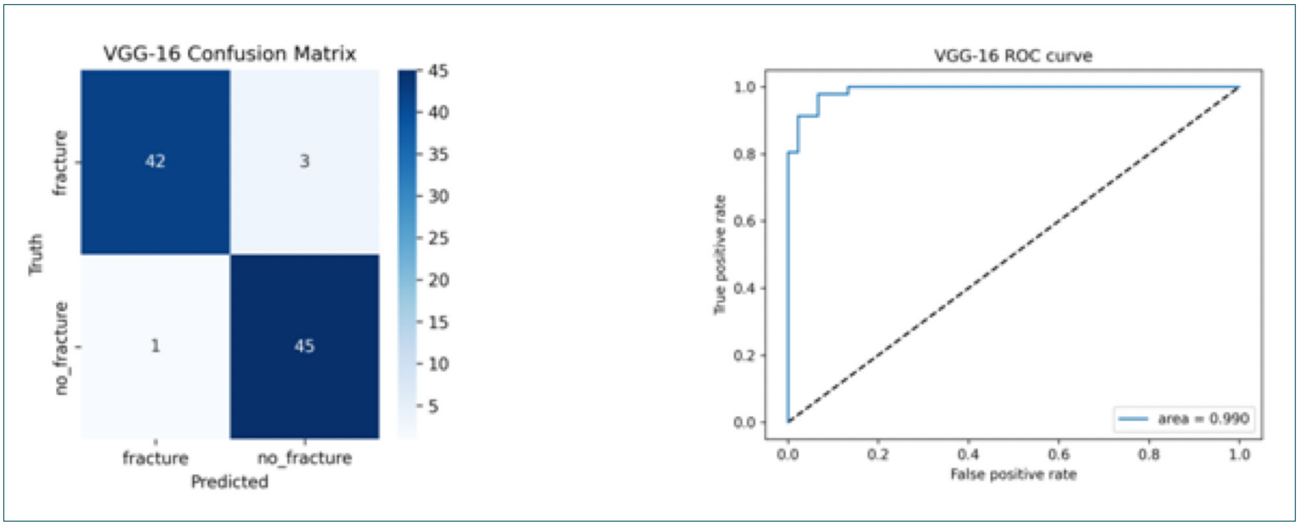

Figure 3. Confusion matrix and receiver operating characteristic (ROC) curve of the VGG-16 model obtained using test data for femoral neck fracture classification.

Table 3. Performance metrics of machine learning algorithms on test data for femoral neck fracture classification

	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	Cohen's Kappa	AUC
k-NN	96.6	93.4	100.0	100.0	96.6	0.93	0.96
LR	95.5	93.4	97.7	97.7	93.4	0.91	0.99
SVM	95.5	93.4	97.7	97.7	95.5	0.91	0.95
RF	93.3	89.1	97.7	97.6	93.1	0.86	0.98
AdaBoost	93.3	86.9	100.0	100.0	93.0	0.86	0.98
NB	92.2	86.9	97.7	97.5	91.9	0.84	0.97
DT	87.7	91.3	84.0	85.7	88.4	0.75	0.86

k-NN: k-Nearest Neighbors; LR: Logistic Regression; SVM: Support Vector Machine; RF: Random Forest; NB: Gaussian Naive Bayes; DT: Decision Tree.

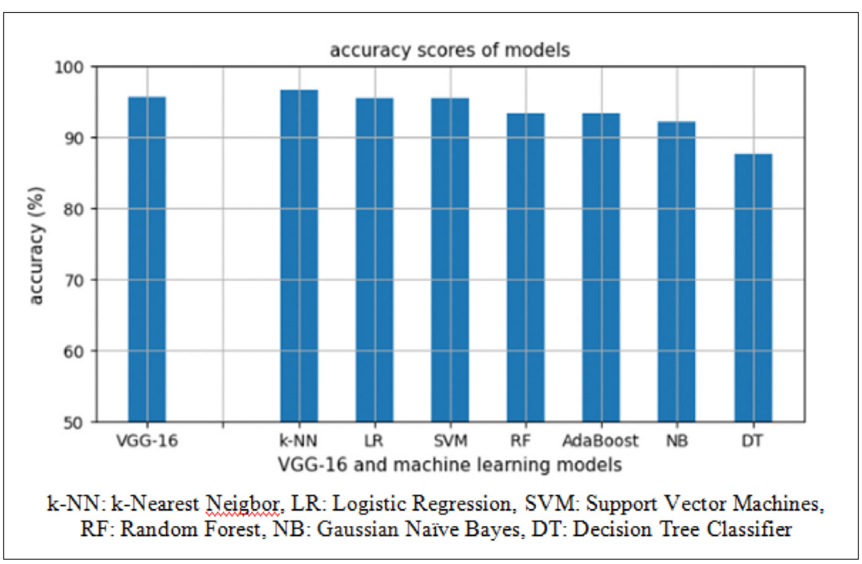

Figure 4. Accuracy scores of the VGG-16 network and machine learning algorithms in femoral neck fracture classification.

ing several algorithms, including logistic regression, support vector machines, k-nearest neighbors (k-NN), decision tree, random forest, Naive Bayes, and AdaBoost. Table 3 shows the performance metrics achieved by each machine learning algorithm. Figure 4 presents a comparison of accuracy scores between the VGG-16 network and the machine learning algorithms for femoral neck fracture classification.

DISCUSSION

To compare deep learning and machine learning algorithms for the detection and classification of femoral neck fractures, transfer learning was first applied using pre-trained VGG-16, ResNet-50, and MobileNetV2 networks. Among these, the pre-trained VGG-16 model outperformed the others, achieving 95.6% accuracy (alpha=0.95, CI: 0.94-0.96), 95.5% sensitivity (alpha=0.95, CI: 0.94-0.96), 93.3% specificity (alpha=0.95, CI: 0.91-0.94), 95.7% precision (alpha=0.95, CI: 0.94-0.97), a 95.5% F1 score (alpha=0.95, CI: 0.94-0.96), a Cohen's kappa of 0.91, and an AUC of 0.99. The VGG-16 network was also used as a feature extractor, and the extracted features were classified using machine learning algorithms. Among these, the k-NN algorithm achieved an accuracy that was 1% higher than that of the VGG-16 model. To our knowledge, this is the first study to use deep learning and machine learning methods for the detection and classification of femoral neck fractures, as well as to compare the performance of these methods.

Cohen's kappa is a measure of agreement used to assess the consistency between two observers, a statistical method that indicates whether the agreement between classifiers is due to chance. It takes values between -1 and +1. A kappa value between 0.81 and 1 indicates very good agreement between classifiers, suggesting that the result is not due to chance. In this study, the Cohen's kappa scores of the deep learning networks were above 0.85. The kappa scores for the k-NN, logistic regression, and support vector machine algorithms was above 0.90. For the random forest, AdaBoost, and Naive Bayes algorithms, the Cohen's kappa scores were above 0.81, indicating perfect agreement between the classifiers.[22]

Several studies have applied deep learning methods for the diagnosis and detection of hip fractures using plain pelvic X-ray images. In these studies, custom CNN architectures or transfer learning with pre-trained networks were used. Beyaz et al. (2020) used a CNN architecture in their study and achieved an accuracy of 83%.[23] Sato et al. (2021) applied a transfer learning method using EfficientNet-B4 in their study and achieved an accuracy of 96.1%.[18] Cheng et al. (2019) used a CNN architecture and reported an accuracy of 91%.[7] Krogue et al. (2020) applied transfer learning with DenseNet-121 and obtained an accuracy of 93.7%.[20] Urakawa et al. (2019) employed a CNN architecture and achieved an accuracy of 95.5%.[24] Gale et al. (2017) used transfer learning with DenseNet and reported an accuracy of 97%.[25]

In our study, transfer learning was first applied. Unlike previous studies, feature extraction was performed using the pre-trained VGG-16 network, and the extracted features were classified using machine learning algorithms. Hyperparameter optimization for the machine learning algorithms was conducted using the grid search method, and better performance metrics were achieved compared to most previous studies.

This study has some limitations. First, the number of pelvic X-rays used was relatively small, with a total of 598 radiographs. In deep learning, model performance generally improves with larger datasets. To address this, data augmentation and transfer learning techniques were employed. Another limitation is that the study was conducted using radiographs from a single center, which limits the generalizability of the model. This limitation could be addressed in future multicenter studies.

Deep learning and machine learning algorithms were used in this study for the detection and classification of femoral neck fractures. Successful results were achieved using pre-trained networks, along with data augmentation and transfer learning techniques. Features automatically extracted from CNN models were utilized for classification tasks by machine learning algorithms. This method can be further improved through multicenter studies and may be particularly valuable in emergency settings, where access to a specialist is limited. The proposed approach can assist physicians working in emergency departments, especially those who do not have sufficient experience in interpreting plain pelvic radiographs. Accurate diagnosis of femoral neck fractures using plain pelvic radiographs may also reduce the need for advanced imaging methods such as MRI or CT.

Ethics Committee Approval: This study was approved by the TC Nevşehir Hacı Bektaş Veli University Ethics Committee (Date: 19.04.2023, Decision No: 2023/04).

Peer-review: Externally peer-reviewed.

Authorship Contributions: Concept: H.F.S., K.Ü.; Design: K.Ü.; Supervision: H.F.S., K.Ü.; Resource: K.Ü., H.F.S., T.K., G.K.G.; Materials: H.F.S., K.Ü.; Data collection and/or processing: K.Ü., T.K., G.K.G.; Analysis and/or interpretation: K.Ü., H.F.S., T.K., G.K.G.; Literature review: K.Ü., H.F.S.; Writing: K.Ü., H.F.S.; Critical review: K.Ü., H.F.S.

Conflict of Interest: None declared.

Financial Disclosure: The author declared that this study has received no financial support.

REFERENCES

1. Cha YH, Ha YC, Yoo JI, Min YS, Lee YK, Koo KH. Effect of causes of surgical delay on early and late mortality in patients with proximal hip fracture. Arch Orthop Trauma Surg 2017;137:625-30.
2. Cha Y, Kim JT, Park CH, Kim JW, Lee SY, Yoo JI. Artificial intelligence and machine learning on diagnosis and classification of hip fracture: Systematic review. J Orthop Surg Res 2022;17:520.
3. Kim KC, Ha YC, Kim TY, Choi JA, Koo KH. Initially missed occult fractures of the proximal femur in elderly patients: Implications for need of operation and their morbidity. Arch Orthop Trauma Surg 2010;130:915-20.

4. Lyons AR. Clinical outcomes and treatment of hip fractures. Am J Med 1997;103:51S-63S.
5. Maheshwari K, Planchard J, You J, Sakr WA, George J, Higuera-Rueda CA, et al. Early surgery confers 1-year mortality benefit in hip-fracture patients. J Orthop Trauma 2018;32:105-10.
6. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29:5469-77.
7. Rehman H, Clement RG, Perks F, White TO. Imaging of occult hip fractures: CT or MRI? Injury 2016;47:1297-301.
8. Saxen A. An introduction to convolutional neural networks. Int J Res Appl Sci Eng Technol 2024;10:943-7.
9. Kim T, Goh TS, Lee JS, Lee JH, Kim H, Jung ID. Transfer learning-based ensemble convolutional neural network for accelerated diagnosis of foot fractures. Phys Eng Sci Med 2023;46:265-77.
10. Nanni L, Paci M, Brahnam S, Lumini A. Comparison of different image data augmentation approaches. J Imaging 2021;7:254.
11. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6:60.
12. Üreten K, Sevinç HF, İğdeli U, Onay A, Maraş Y. Use of deep learning methods for hand fracture detection from plain hand radiographs. Ulus Travma Acil Cerrahi Derg 2022;28:196-201.
13. Huang ST, Huang MY, Liu LR, Tsai MF, Chiu HW. Recognition of rib fracture on chest X-ray images with deep learning: A pilot study [Preprint]. Res Sq 2022.
14. Üreten K, Maraş Y, Duran S, Gök K. Deep learning methods in the diagnosis of sacroiliitis from plain pelvic radiographs. Mod Rheumatol 2023;33:202-6.
15. Duran S, Üreten K, Maraş Y, Maraş HH, Gök K, Atalar E, et al. Automatic detection of spina bifida occulta with deep learning methods from plain pelvic radiographs. Res Biomed Eng 2023;39:655–61.
16. Atalar E, Üreten K, Kanatlı U, Çiçeklidağ M, Kaya İ, Vural A, et al. The diagnosis of femoroacetabular impingement can be made on pelvis radiographs using deep learning methods. Jt Dis Relat Surg 2023;34:298-304.
17. Sato Y, Takegami Y, Asamoto T, Ono Y, Hidetoshi T, Goto R, et al. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: A multicenter study. BMC Musculoskelet Disord 2021;22:407.
18. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ. Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. J Digit Imaging 2020;33:1209-17.
19. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, et al. Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2020;2:e190023.
20. Benyahia S, Meftah B, Lézoray O. Multi-features extraction based on deep learning for skin lesion classification. Tissue Cell 2022;74:101701.
21. McHugh ML. Interrater reliability: The kappa statistic. Biochem Med 2012;22:276-82.
22. Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg 2020;31:175-83.
23. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 2019;48:239-44.
24. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv [Preprint]. 2017. arXiv:1711.06504.
25. Parker MJ. Missed hip fractures. Emerg Med J 1992;9:23–7.

ORİJİNAL ÇALIŞMA - *ÖZ*

Derin öğrenme ve makine öğrenmesi yöntemleriyle düz pelvis röntgenlerinden femur boyun kırıklarının tespiti ve sınıflandırılması

AMAÇ: Femur boyun kırıkları özellikle yaşlılarda ciddi bir sağlık sorunudur. Bu çalışmanın amacı, düz pelvis röntgenlerinden femur boyun kırıklarını derin öğrenme ve makine öğrenmesi algoritmalarıyla teşhis etmek, sınıflandırmak ve bu yöntemlerin performanslarını karşılaştırmaktır.

GEREÇ VE YÖNTEM: Çalışma, femur boyun kırığı olan 296 hastaya ve femur boyun kırığı olmayan 302 kişiye ait toplam 598 düz pelvis röntgen görüntüsü üzerinde yürütülmüştür. Öncelikle, önceden eğitilmiş derin öğrenme modelleri olan VGG-16, ResNet-50 ve MobileNetv2 ağları ile transfer öğrenme yöntemi uygulandı.

BULGULAR: Femur boyun kırığı tespiti ve sınıflandırması için önceden eğitilmiş VGG-16 ağı, ResNet-50 ve MobileNetv2 ağlarına göre biraz daha iyi performans göstermiştir. VGG-16 ağı ile %95.6 doğruluk, %95.5 duyarlılık, %93,3 özgüllük, %95.7 kesinlik, %95.5 F1 skoru, 0.91 Cohen's kappa ve 0.99 roccurve sonuçları elde edilmiştir. Daha sonra, VGG-16 evrişim katmanından çıkarılan öznitelikler, yaygın olarak kullanılan makine öğrenmesi algoritmaları ile yeniden sınıflandırıldı. K-en yakın komşuluk (k-NN) algoritması, diğer makine öğrenmesi algoritmalarından daha iyi performans göstermiştir, k-NN, doğruluk açısından VGG-16 modelinden %1 daha iyi performans göstermiştir.

SONUÇ: Femur boyun kırıklarının tespiti ve sınıflandırılmasında derin öğrenme ve makine öğrenmesi yöntemleri ile başarılı sonuçlar elde edilmiştir. Model, çok merkezli çalışmalarla geliştirilebilir. Önerilen model, özellikle acil servislerde çalışan hekimler ve düz pelvis radyografilerini değerlendirmede yeterli deneyime sahip olmayan hekimler için yararlı olabilir.

Anahtar sözcükler: Derin öğrenme; femur boyun kırıkları; makine öğrenmesi; önceden eğitilmiş ağlar; özellik çıkarma.

Ulus Travma Acil Cerrahi Derg 2025;31(8):783-788 DOI: 10.14744/tjtes.2025.75806