Joint Diseases and Related Surgery

ORIGINAL ARTICLE

Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches

Salih Beyaz, MD[1], Koray Açıcı, MD[2], Emre Sümer, MD[2]

[1]Department of Orthopedics and Traumatology, Başkent University Adana Turgut Noyan Training and Research Centre. Adana, Turkey
[2]Department of Computer Engineering, Başkent University, Ankara, Turkey

Femoral fractures are a major health issue faced by the elderly population.[1,2] Incidence is predicted to double in the next 30 years, in parallel with the aging global population.[3-5] Early diagnosis and treatment not only facilitate the protection of the joint, but also help patients to sustain their quality of life and ambulation capacity in the postoperative period.[6] Pelvic X ray (PXR) is the simplest, cheapest, and fastest method for the diagnosis of femoral fractures. However, it does not provide 100% accuracy (Acc). It has been reported that about 2% all of hip fractures are not diagnosed by simple PXR.[7,8] Misdiagnosis in turn causes late treatment, extended postoperative recovery time, and increased treatment costs. Although femoral neck fracture detection rates using magnetic resonance imaging (MRI), computed tomography (CT), and radionuclide methods are higher, their routine use is not cost-effective.[9]

The use of deep learning techniques in the field of medical image processing has increased in popularity in recent years. The facts that computers are not affected by environmental factors, do not forget what they have learned, and possess unlimited memory capacity suggest that if Acc rates are improved,

Received: November 05, 2019
Accepted: January 02, 2020
Published online: March 26, 2020

Correspondence: Salih Beyaz, MD. Başkent Üniversitesi Adana Turgut Noyan Eğitim ve Araştırma Merkezi Ortopedi ve Travmatoloji Kliniği, 01240 Yüreğir, Adana, Türkiye.

E-mail: salihbyz@gmail.com

Doi: 10.5606/ehc.2020.72163

Citation: 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(2):175-183.

ABSTRACT

Objectives: This study aims to detect frontal pelvic radiograph femoral neck fracture using deep learning techniques.

Patients and methods: This retrospective study was conducted between January 2013 and January 2018. A total of 234 frontal pelvic X-ray images collected from 65 patients (32 males, 33 females; mean age 74.9 years; range, 33 to 89 years) were augmented to 2106 images to achieve a satisfactory dataset. A total of 1,341 images were fractured femoral necks while 765 were non-fractured ones. The proposed convolutional neural network (CNN) architecture contained five blocks, each containing a convolutional layer, batch normalization layer, rectified linear unit, and maximum pooling layer. After the last block, a dropout layer existed with a probability of 0.5. The last three layers of the architecture were a fully connected layer of two classes, a softmax layer and a classification layer that computes cross entropy loss. The training process was terminated after 50 epochs and an Adam Optimizer was used. Learning rate was dropped by a factor of 0.5 on every five epochs. To reduce overfitting, regularization term was added to the weights of the loss function. The training process was repeated for pixel sizes 50×50, 100×100, 200×200, and 400×400. The genetic algorithm (GA) approach was employed to optimize the hyperparameters of the CNN architecture and to minimize the error after testing the model created by the CNN architecture in the training phase.

Results: Performance in terms of sensitivity, specificity, accuracy, F1 score, and Cohen's kappa coefficient were evaluated using five-fold cross validation tests. Best performance was obtained when cropped images were rescaled to 50×50 pixels. The kappa metric showed more reliable classifier performance when 50×50 pixels image size was used to feed the CNN. The classifier performance was more reliable according to other image sizes. Sensitivity and specificity rates were computed to be 83% and 73%, respectively. With the inclusion of the GA, this rate increased by 1.6%. The detection rate of fractured bones was found to be 83%. A kappa coefficient of 55% was obtained, indicating an acceptable agreement.

Conclusion: This experimental study utilized deep learning techniques in the detection of bone fractures in radiography. Although the dataset was unbalanced, the results can be considered promising. It was observed that use of smaller image size decreases computational cost and provides better results according to evaluation metrics.

Keywords: Artificial intelligence; deep learning, femoral neck fracture, genetic algorithm.

algorithms will directly influence the decision-making processes of doctors regarding their patients in the near future.

The benefits of using deep learning techniques are early diagnosis of fractures, early initiation of treatment, shortening of postoperative recovery time, and prevention of increased costs due to misdiagnosis. The present study differs from others in its use of genetic algorithms (GAs) in addition to deep learning algorithm. Thus in this study, we aimed to detect PXR femoral neck fracture using deep learning techniques.

PATIENTS AND METHODS

This retrospective study was conducted at Başkent University Adana Turgut Noyan Training and Research Centre between January 2013 and January 2018. A total of 234 proximal PXR images were collected from 65 patients (32 males, 33 females; mean age 74.9 years; range, 33 to 89 years). Fractures were observed in 149 images while no fractures were observed in the remaining 85. Data augmentation methods were used to increase the size of the dataset.[10] To do so, the original images were rotated by 10, 20, and 30 degrees in clockwise and counterclockwise directions. In addition, Gaussian noise was added and mirror images were obtained. As a result, the total dataset included 2,106 images, of which 1,341 were fractured femoral necks and 765 were non-fractured ones. The synthetically generated images obtained by data augmentation methods are shown in Figure 1. The study protocol was approved by the Başkent University Ethics Committee (Project no: KA19/46). The study was conducted in accordance with the principles of the Declaration of Helsinki.

Convolutional neural network (CNN) architecture is a deep learning approach commonly used in theoretical and practical studies on topics such as disease classification, MRI reconstruction, and effector protein prediction.[11-13] A common CNN architecture consists of input, convolutional, pooling, and fully connected layers. Additionally, batch normalization, rectified linear unit, and dropout layers can also be used to speed up the process and reduce overfitting. In the convolution layer, by using convolution operation, the input image is filtered resulting in feature maps. Basic features, basic patterns, and advanced patterns can be discovered from these feature maps depending on the number of convolutional layers used. The pooling layer reduces the size of the feature maps generated by the convolutional layer. As the size of the feature maps decrease, the elapsed time of the training phase is reduced. The fully connected layer converts the resulting feature maps into a pixel vector to prepare the input matrix for classification. In the current study, CNN architecture was used to classify the fractured and non-fractured bones.

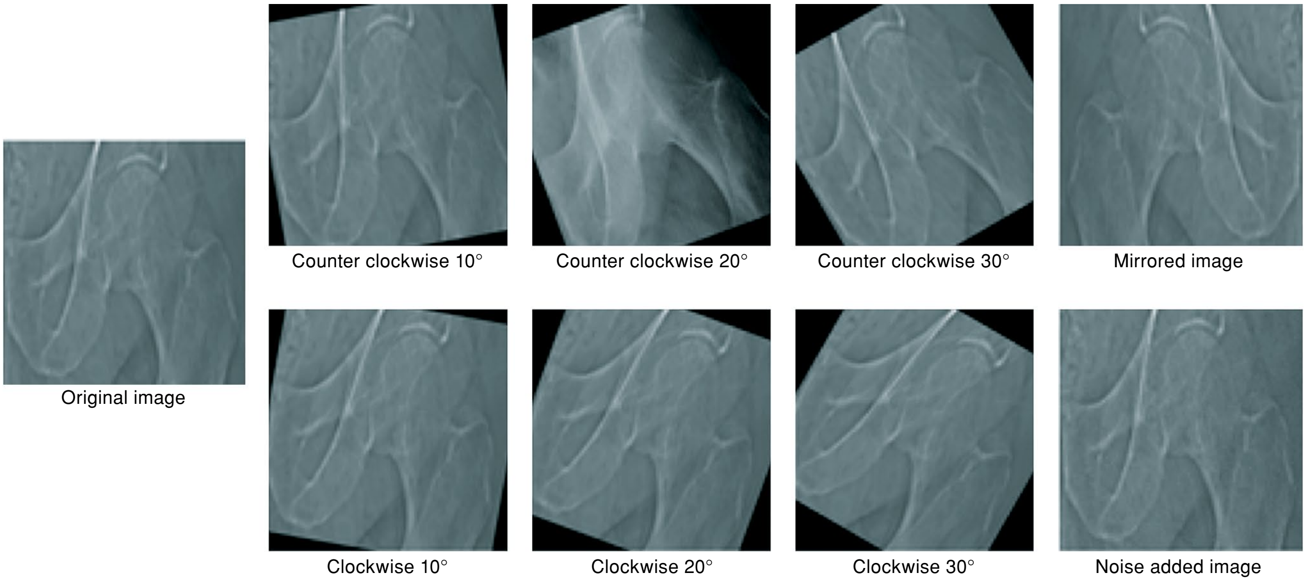


FIGURE 1. Sample images from augmented dataset.

Genetic algorithms can be described as a heuristic search that simulates the evolutionary process and are considered as a probabilistic optimization method based on the principles of evolution. Genetic algorithms are widely used in search and optimization problems to optimize user interface layouts, cancer diagnosis, and medical image denoising.[14-16]

When using GAs, each possible solution is accepted as an individual or a chromosome in the search space of the problem. In GA, the solution or individual refers to the values of the parameters to be optimized while the search space refers to the boundaries of the parameters. Individuals form the population and each individual is composed of genes which are the parameters needed to be optimized in the problem. The GA tries to find the optimal individual that satisfies the fitness function. In this study, GA was used to optimize classification performance.

The pseudo code of the GA is given below:

1. Initialize population randomly.
2. Evaluate fitness value of individuals in the population.
3. When stopping criteria are not met:
   - Choose individuals for next generation.
   - Apply crossover and mutation.
   - Evaluate again.

The workflow of the proposed framework is illustrated in Figure 2. Regions containing both fractured and non-fractured femoral necks were cropped from the X-ray images manually. These cropped images were than rescaled to a specific size. To provide additional data to the CNN in the training phase, data augmentation was applied to all rescaled images. All cropped images became the same size, training with the CNN was initiated, and a model was constructed. Finally, test images were classified according to the constructed model as "fractured" or "non-fractured". A GA block was used to optimize the hyperparameters of the CNN.

The CNN architecture in the proposed framework had five blocks, each including a convolutional layer, batch normalization layer, a rectified linear unit, and maximum pooling layer. Each convolutional layer had 3×3 filters that produced 8, 16, 32, 32, and 32 feature maps, respectively. Batch normalization layers in each block took part between a convolutional layer and a rectified linear unit and normalized the activations for speeding up the training. Rectified linear unit applied a threshold to the input. Maximum pooling layers reduced the size of the feature maps using a filter size of 2×2 and a stride value of two. After the last block, a dropout layer existed with a probability of 0.5. The last three layers of the CNN architecture were a fully connected layer that had two outputs (classes), a softmax layer, and a classification layer that computed cross entropy loss.

In the training process, the AdamOptimizer was utilized. The process was terminated after 50 epochs. Before each epoch, the training data were shuffled.

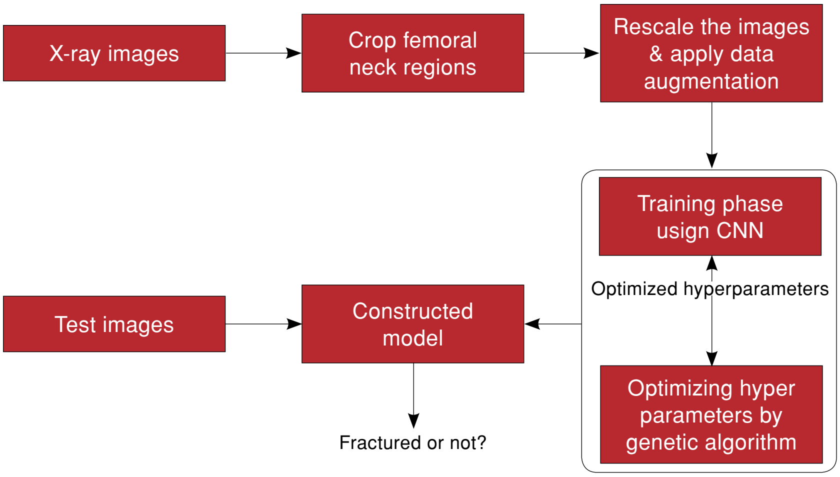

X-ray images
Crop femoral neck regions
Rescale the images & apply data augmentation
Training phase usign CNN
Optimized hyperparameters
Optimizing hyper parameters by genetic algorithm
Test images
Constructed model
Fractured or not?

FIGURE 2. General framework.
CNN: Convolutional neural network.

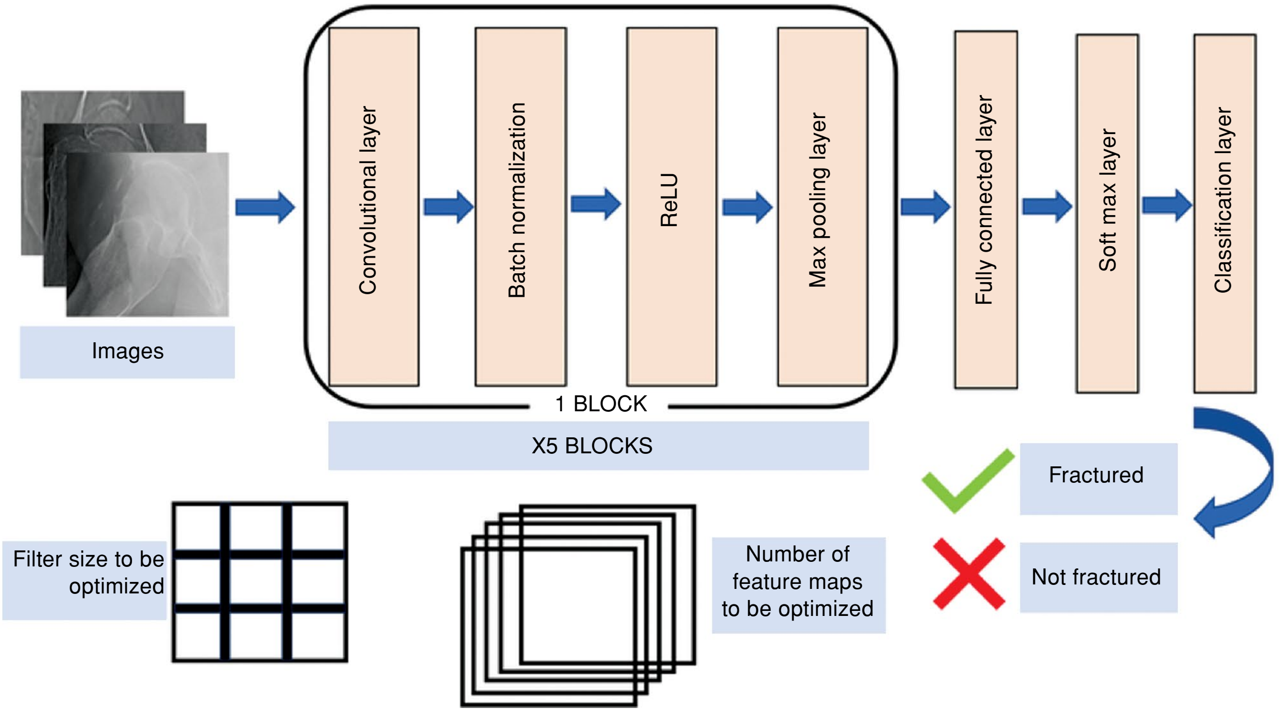


FIGURE 3. Proposed convolutional neural network architecture.

Learning rate was dropped by a factor of 0.5 on every five epochs. Our CNN architecture is given in Figure 3. The training process was repeated for four image sizes of 50×50, 100×100, 200×200, and 400×400 pixels.

The GA had population sizes of 10, 50, and 100 chromosomes for three different setups. The algorithm was repeated until 10 generations were reached. One chromosome had 10 variables or genes. The first five represented the filter sizes and the second five represented the number of feature maps produced by the convolutional layers. Since there were five convolutional layers in the CNN architecture, the filter size and feature maps of each layer should have been optimized. A representation of a chromosome is given in Figure 4.

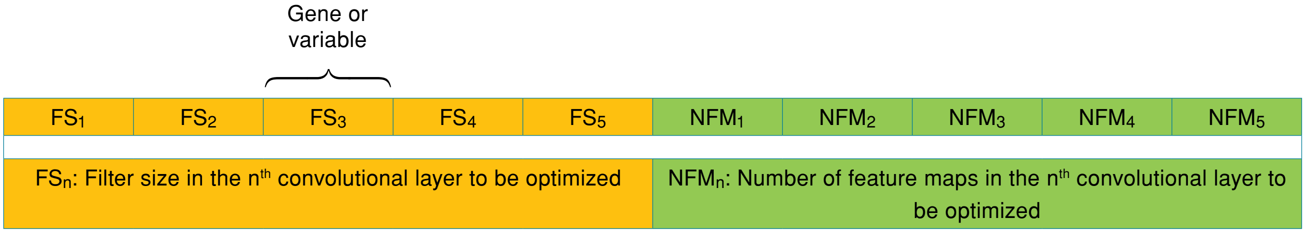
To select parents for crossover, stochastic uniform selection was used. The selection algorithm arranged each candidate parent on a line according to its length. This length was calculated by dividing the fitness value by the sum of all fitness values. Each candidate occupied a space according to its probability of being selected on the line. A random number and a step size were generated for initiating the selection. The algorithm used a random number to choose the first parent, and then moved along the line by the step size to choose other parents.

After selection of the parents, a random binary vector was created for the crossover operation. Since the random binary vector contained only zeros and ones, its bits represented the first parent (for one) and second parent (for zero) to create a child. This algorithm was labeled as the scattered crossover and is presented in Figure 5. The mutation operation simply changed the genes in the chromosomes randomly.

Five-fold cross validation tests were performed to obtain the performance of the proposed framework in terms of sensitivity *(i)*, specificity *(ii)*, Acc *(iii)*, F1 score *(iv)*, and Cohen's kappa coefficient *(v)*. Five metrics for classification performance are given below. Referred to true positives-TP (fractured femoral neck), false positives-FP (fractured incorrectly), true negatives-TN (non-fractured femoral neck), and false negatives-FN (non fractured incorrectly), respectively.

Sensitivity (Sn) was measured as the ratio between correctly classified fractured femoral necks and all fractured femoral necks:

$$(i)\ Sn=TP/(TP+FN)$$


Gene or variable

FS_1	FS_2	FS_3	FS_4	FS_5	NFM_1	NFM_2	NFM_3	NFM_4	NFM_5

FS_n: Filter size in the n^{th} convolutional layer to be optimized

NFM_n: Number of feature maps in the n^{th} convolutional layer to be optimized

FIGURE 4. Chromosome representation in genetic algorithm.

Parent 1	a	b	c	d	e	f	g	h	i	j
Parent 2	k	l	m	n	o	p	q	r	s	t
Random binary vector	1	1	0	0	1	0	0	1	0	0
Child	a	b	m	n	e	p	q	h	s	t

FIGURE 5. Crossover operation in genetic algorithm.

Specificity (Sp) was measured as the ratio between correctly classified non-fractured femoral necks and all non-fractured femoral necks:

$$(ii)\ Sp=TN/(TN+FP)$$

Accuracy was the ratio between correctly classified non-fractured & fractured femoral necks and all samples:

$$(iii)\ Acc=(TP+TN)/(TP+FP+TN+FN)$$

F1 score (F1) was measured as the harmonic average of the precision and recall and used as a measure of a test's Acc:

$$(iv)\ F_1=(Precision*Recall)/(Precision+Recall)*2$$

$$Precision=TP/(TP+FP)$$

$$Recall=TP/(TP+FN)$$

Cohen's kappa (kappa) coefficient was used to measure inter-rater agreement for categorical items by considering the chance factor of agreement:

$$(v)\ Kappa=(Acc-p_e)/(1-p_e)$$

$$p_e=(TN+FP)*(TN+FN)+(TP+FN)*(TP+FP)/(TN+TP+FN+FP)^2$$

RESULTS

The results of the five-fold cross validation experiments were obtained using the evaluation metrics (Table I). The best performance in terms of Acc, kappa, F1 score, and specificity was obtained when cropped images were rescaled to 50×50 pixels. The best sensitivity result was obtained when 100×100 rescaled images were used. On the other hand, kappa metric showed that when 50×50 pixels image size was used to feed the CNN, the classifier performance was more reliable than the other image sizes. The evaluation results and optimized hyperparameters using GA are given in Table II.

Selection of a population size of 10 resulted in a 1% decrease (from 77.7 to 76.7%) in Acc for images of 50×50 pixels. The best results in terms of Acc were reached by using a population of 50 and 100. Although a population size of 100 outperformed using a population size of 50 in terms of sensitivity, the reverse was determined in terms of specificity. F1 scores were very close to the Acc. The kappa coefficient showed that the reliability of the classifier

TABLE I
Performance comparison for different image sizes

Image size	Accuracy	Kappa	F_1	Sensitivity	Specificity
50×50	0.777	0.518	0.825	0.825	0.693
100×100	0.770	0.497	0.823	0.837	0.654
200×200	0.729	0.394	0.796	0.830	0.552
400×400	0.712	0.389	0.780	0.803	0.552

TABLE II
Performance comparison for different population sizes using genetic algorithm on 50×50 image size

Population size	Optimized hyperparameters Filter size Number of feature maps	Accuracy	Kappa	F_1	Sensitivity	Specificity
10	4 10 4 9 23 88 54 81 69 118	0.767	0.497	0.817	0.817	0.681
50	3 4 10 10 15 18 122 117 80 79	0.793	0.554	0.836	0.829	0.729
100	3 10 24 9 24 111 123 114 56 73	0.793	0.552	0.838	0.838	0.714

performances was almost the same when using a population size of 50 or 100. On the other hand, using a small population size decreased computational time and thus a population size of 50 can be accepted as the optimal solution in the present case. Values varied for both filter sizes and number of feature maps (Table II). As a result, the use of GA improved overall Acc by 1.6%. For other metrics, the effect of using GA on performance can be seen in Figure 6.

Additionally, confusion matrices for two different scenarios (with and without GA) are given in Figure 7. According to FN statistics, of 1,341 fractured images, the system mis-detected 235

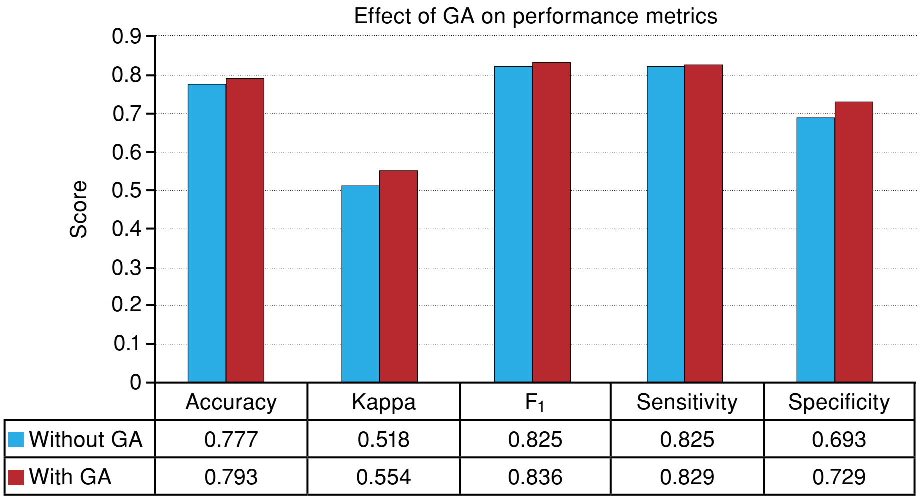


FIGURE 6. Effect of genetic algorithm on performance metrics.
GA: Genetic algorithm.

True negatives	False negatives
530	235
False positives	True positives
235	1,106

(a) Confusion matrix for 50×50 image size without GA

True negatives	False negatives
558	230
False positives	True positives
207	1,111

(b) Confusion matrix for 50×50 image size with GA and population size of 50

FIGURE 7. Confusion matrices with and without genetic algorithm.
GA: Genetic algorithm.

images versus 230 with GA included. False positive rates revealed that of the 765 non-fractured images, 235 were labeled as fractured if we excluded GA. On the other hand, this number was reduced to 207 if we included GA. In general, the system successfully detected approximately 83% of fractured images. Similarly, the detection of non-fractured cases was computed at approximately 73%.

DISCUSSION

In this study, femoral neck fracture was detected in the 50×50 size anteroposterior (AP)-pelvis radiography with 83% Acc using the GA-supported deep learning approach. To our knowledge, this is the first study in which femoral neck fractures were detected using deep learning techniques. Despite the modest sample size and unstable data distribution, sensitivity and specificity values of 83% and 73% were obtained, respectively.

Extraordinary developments have occurred in the image processing technology over the last two decades. It is anticipated that the use of this technology in the processing and evaluation of medical images will facilitate the diagnosis of diseases and regulation of treatments. Olczak et al.[17] worked on distal radius radiographies due to the high incidence rate of these types of fractures and calculated a fracture detection rate in the distal radius radiography of 83%.

Kim and MacKinnon[18] tested deep learning algorithms, which they trained with 694 non-fracture wrist radiographies against 695 distal radius fracture radiographies and with 50 distal radius fracture and 50 non-fracture distal radius plain radiographies. They reported sensitivity of 0.9 and specificity of 0.88. Additionally, they claimed that this technique is largely transferable, and therefore, has many potential applications in medical imaging, which may lead to significant improvements in workflow productivity and clinical risk reduction.

Furthermore, Pranata et al.[19] obtained 98% sensitivity in the classification of calcaneus fractures with the help of CT images and deep learning technique. In their study, two types of CNN architectures with different network depths were evaluated and compared for the classification performance of CT scans into fracture and non-fracture categories based on coronal, sagittal, and transverse views. Their bone fracture detection algorithm used the speeded-up robust features method, Canny edge detection, and contour tracing to incorporate fracture area matching. They were very successful in both the detection and classification of the fraction. Compared to conventional, CT scan has higher resolution and better image quality. However, patients are exposed to a higher dose of radiation during CT scan.

Unlike other studies, Cheng et al.[20] used the Deep Convolutional Neural Network (DCNN) method to detect hip fractures in AP plain pelvis radiography. They trained the system they developed with 25,505 PXR. The authors also used visualization algorithm gradient-weighted class activation mapping to confirm the validity of the model. Algorithm Acc was reported as 91% and sensitivity as 98% while the false-negative rate for identifying hip fractures was 2%, and area under the curve (AUC) was 0.98. In addition, 95.9% Acc was determined in lesion identification using the visualization algorithm.[20] Cropped images are utilized in deep learning techniques that evaluate medical images to increase validation Acc and to avoid "black box" mechanisms.[21,22] When cropped to include the features required for target recognition, DCNN will detect the lesion more easily and quickly. Cheng et al.[20] reduced the image matrix size to 512×512 pixels instead of cropping images. The study's most important development was that conclusion was reached without cropping the radiological image and marking a specific area.

Gale et al.[21] trained the algorithm they developed to detect hip fractures using 53,000 AP pelvis radiographies. During this training, they reduced the image resolution of the PXR from 3000×3000 pixels to 1024×1024. Additionally, patients who previously underwent hip surgery were trained separately by a radiologist. They found that an AUC value of 0.994 was, to the best of their knowledge, the highest level ever reported for automated diagnosis in any large scale medical task, not just in radiology. They stated that this detection rate was at a similar level to expert radiologist.

Hemanth and Anitha[23] proposed modified GAs for feature selection. Their aim was to reduce the number of features that were obtained from abnormal magnetic resonance brain tumor images to feed the back propagation neural network classifier. As a result, their proposed method yielded 98% Acc, 96% sensitivity, and 98% specificity with reduced number of features.

Although the results were found to be promising, there are some limitations of our study. First of all, we worked on a small-sized dataset. Since deep learning methods require as many images as possible, we tried to overcome this situation by applying data augmentation techniques. By doing so, we have shown that promising results can be achieved with a small-sized dataset. However, increasing the number of X-ray images still continues to be an important factor for achieving better results. In addition to increasing the number of images, obtaining more images from different healthcare facilities will increase the diversity and more reliable results can be achieved. All experiments in this study were run on an outdated graphic card. With the use of new generation and multiple graphics cards, training time will be further reduced and the number of experiments with different setups will increase. As a future work, we aim to include MRI scans in addition to X-ray images. Heuristic approaches such as ant colony, simulated annealing, and particle swarm optimization are planned to be compared with GA for optimization.

In conclusion this experimental study utilized CNNs in the detection of bone fractures in radiography. The trained model yielded an overall Acc of 77.7% when 50×50 image sizes were used. With the inclusion of GA, this rate increased by 1.6%. The detection rate of fractured bones was found to be 83%. A kappa coefficient of 55% was obtained, indicating an acceptable agreement. Although the dataset was unbalanced, the results can be considered promising. It was observed that use of smaller image size decreases computational cost and provides better results according to evaluation metrics.

Declaration of conflicting interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

This study was supported by Başkent University Research Fund.

REFERENCES

1. Atik OŞ. There is an association between sarcopenia, osteoporosis, and the risk of hip fracture. Eklem Hastalik Cerrahisi 2019;30:1.
2. Bozkurt HH, Tokgöz MA, Yapar A, Atik OŞ. What is the importance of canal-to-diaphysis ratio on osteoporosis-related hip fractures? Eklem Hastalik Cerrahisi 2019;30:296-300.
3. Leslie WD, O'Donnell S, Jean S, Lagacé C, Walsh P, Bancej C, et al. Trends in hip fracture rates in Canada. JAMA 2009;302:883-9.
4. Lewiecki EM, Wright NC, Curtis JR, Siris E, Gagel RF, Saag KG, et al. Hip fracture trends in the United States, 2002 to 2015. Osteoporos Int 2018;29:717-22.
5. Bozkurt HH, Atik OŞ, Tokgöz MA. Can distal radius or vertebra fractures due to low-energy trauma be a harbinger of a hip fracture? Eklem Hastalik Cerrahisi 2018;29:100-3.
6. Grimes JP, Gregory PM, Noveck H, Butler MS, Carson JL. The effects of time-to-surgery on mortality and morbidity in patients following hip fracture. Am J Med 2002;112:702-9.
7. Dominguez S, Liu P, Roberts C, Mandell M, Richman PB. Prevalence of traumatic hip and pelvic fractures in patients with suspected hip fracture and negative initial standard radiographs--a study of emergency department patients. Acad Emerg Med 2005;12:366-9.
8. Perron AD, Miller MD, Brady WJ. Orthopedic pitfalls in the ED: radiographically occult hip fracture. Am J Emerg Med 2002;20:234-7.
9. Rehman H, Clement RG, Perks F, White TO. Imaging of occult hip fractures: CT or MRI? Injury 2016;47:1297-301.
10. Dao T, Gu A, Ratner AJ, Smith V, De Sa C, Ré C. A Kernel Theory of Modern Data Augmentation. Proc Mach Learn Res 2019;97:1528-37.
11. Sengur A, Akbulut Y, Guo Y, Bajaj V. Classification of amyotrophic lateral sclerosis disease based on convolutional neural network and reinforcement sample learning algorithm. Health Inf Sci Syst 2017;5:9.
12. Wu Y, Ma Y, Liu J, Du J, Xing L. Self-attention convolutional neural network for improved MR image reconstruction. Information Sciences 2019;490:317-28.
13. Açıcı K, Aşuroğlu T, Erdaş ÇB, Oğul H. T4SS effector protein prediction with deep learning. Data 2019,4,45.
14. Diego-Mas JA, Garzon-Leal D, Poveda-Bautista R, Alcaide-Marzal J. User-interfaces layout optimization using eye-tracking, mouse movements and genetic algorithms. Appl Ergon 2019;78:197-209.
15. Gorunescu F, Belgiuc S. Genetic algorithms for breast cancer diagnostics. Encyclopedia of Biomedical Engineering 2019;380-8.
16. Liu P, El Basha MD, Li Y, Xiao Y, Sanelli PC, Fang R. Deep Evolutionary Networks with Expedited Genetic Algorithms for Medical Image Denoising. Med Image Anal 2019;54:306-15.
17. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581-6.
18. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73:439-45.
19. Pranata YD, Wang KC, Wang JC, Idram I, Lai JY, Liu JW, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. Comput Methods Programs Biomed 2019;171:27-37.
20. 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.

21. Gale W, Rayner LO, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv 2017;1711.06504.

22. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv 2018;1608.06993.

23. Hemanth DJ, Anitha J. Modified Genetic Algorithm approaches for classification of abnormal Magnetic Resonance Brain tumour images. Applied Soft Computing 2019; 75:21-8.