ORIGINAL ARTICLE

Rheumatology Quarterly

DOI: 10.4274/qrheumatol.galenos.2024.85057
Rheumatology Quarterly 2024;2(3):121-9

USING ENSEMBLE LEARNING AND FEATURE SELECTION IN THE DIAGNOSIS OF LOW BACK PAIN

Yüksel Maraş[1], Ahmet Kor[2], Semra Duran[3], Kevser Orhan[4], Ebru Atalar[4], Emine Sena Sözen[4], Kemal Üreten[5], Hadi Hakan Maraş[6]

[1]University of Health Sciences Turkey, Ankara Bilkent City Hospital, Clinic of Rheumatology, Ankara, Turkey
[2]Aksaray Training and Research Hospital, Clinic of Rheumatology, Aksaray, Turkey
[3]University of Health Sciences Turkey, Ankara Bilkent City Hospital, Clinic of Radiology, Ankara, Turkey
[4]University of Health Sciences Turkey, Ankara Bilkent City Hospital, Clinic of Rheumatology, Ankara, Turkey
[5]Çankaya University Faculty of Engineering, Department of Computer Engineering, Ankara, Turkey
[6]Çankaya University Vocational School, Department of Computer Programming, Ankara, Turkey

Abstract

Aim: A wide variety of research is currently being conducted on how artificial intelligence can assist clinical decision-making and improve clinician judgments. The goal of this research was to develop a computer-aided diagnostic (CAD) approach that can aid healthcare professionals in identifying lumbosacral pathologies.

Material and Methods: The study included 633 abnormal and 442 normal lateral lumbosacral radiographs, and the You Only Look Once algorithm was used to automate the cropping task. This study used pre-trained VGG-16, ResNet-101, and MobileNetV2 models for transfer learning. Feature extraction was performed from the intermediate layer of VGG-16, resulting in 512 features. Then, a variance threshold was applied, resulting in 221 selected features with a variance threshold of 0.01. Then, support vector classifier, logistic regression, random forest classifier, and k-nearest neighbours machine learning models were trained using both sets of 512 extracted features and 221 selected features separately.

Results: The results from the ensemble learning model with the stacking classifier using features selected using a threshold value 0.01 from features extracted were: accuracy 93.0% (best); sensitivity, 91.8%; specificity, 94.1%; precision, 92.9%; F1 score, 92.3% (best); area under the receiver operating characteristic curve, 0.97 (one of the best); and Cohen's kappa, 0.86 (best).

Conclusion: The ensemble learning model with a stacking classifier using features selected by using a threshold value of 0.01 from features extracted by processing the intermediate layer of VGG-16 performs better than the transfer learning models using pre-trained networks, such as VGG-16, ResNet-50, and MobileNetV2, and the learning methods that do not apply feature selection in distinguishing lumbar vertebral pathologies.

Keywords: Artificial intelligence, computer-aided diagnosis, convolutional neural networks, deep learning, low back pain, machine learning, ensemble learning, feature selection

Address for Correspondence: Yüksel Maraş, University of Health Sciences Turkey, Ankara Bilkent City Hospital, Clinic of Rheumatology, Ankara, Turkey
Phone: +90 555 332 97 40 **E-mail:** ymaras@hotmail.com **ORCID ID:** orcid.org/0000-0001-9319-0955
Received: 09.08.2024 **Accepted:** 24.09.2024

INTRODUCTION

Chronic low back pain (lumbar spine pain) is a widespread problem that most people experience at some point in their lives (1). It is rarely fatal, and it is usually benign and self-limiting. The first step in managing lumbar low back pain is plain radiography, which often provides an anteroposterior and lateral view of the lumbar vertebrae (2). The most common causes of low back pain on plain radiography include disc space narrowing, osteophytes, spondylosis, endplate sclerosis, spondylolisthesis, and facet joint osteoarthritis (3).

The prevalence of low back pain is increasing and can result in decreased physical function. Plain radiographs are widely used in clinical practice because they are relatively inexpensive, are easy to apply, and have become standard for patients with low back pain (1,3).

The ordering clinician primarily evaluates plain radiographs of the lumbar spine. For this reason, computer-assisted diagnosis as a primary aid is becoming more common (4). Studies have shown that the success of diagnosis by clinicians can be increased through the use of a deep learning method (5-9).

Deep learning methods, a subset of machine learning, are effective in extracting complex features from raw data such as images, text, and audio. Convolutional neural networks (CNNs), which use a deep learning architecture, consist of many layers and have succeeded in image processing to retrieve information from images. Different CNN architectures have been developed for different purposes, such as classification, segmentation, object detection, and localization (10,11). Due to these features, deep learning methods have been successfully used in applications such as object recognition, speech recognition, face recognition, text analysis, language modeling, translation, autonomous vehicles, robotic applications, e-commerce, medical image analysis, disease diagnosis, and treatment planning. However, large amounts of data are required to implement CNNs for image classification, and it is difficult to find sufficient data in the medical field. Networks pre-trained on the Imagenet dataset consisting of natural images were successfully used to classify medical images using the transfer learning method (12-15). These pre-trained networks included AlexNet (16), GoogLeNet (17), VGG (18), ResNet (10), MobileNet (19), and DenseNet (20).

The region-based CNN (RCNN), Fast RCNN, Faster RCNN, and YOLO CNN architectures are used for segmentation and object detection (21,22). YOLO is fast and has been successfully applied to object detection tasks. It treats object detection as a regression problem and performs real-time object detection (45 frames per second) using a single CNN called DarkNet. The proposed YOLO model makes predictions for various bounding boxes of different sizes and aspect ratios to detect objects of diverse shapes and sizes. The non-max suppression algorithm selects the best option from the multiple projected bounding boxes (23). YOLO works with low-resolution images, and the algorithm is not very successful in detecting small objects; thus, other CNN structures are preferred for classification tasks.

Feature selection is the process of selecting informative and relevant features from a more extensive dataset that better characterizes multiple class patterns (24). Variance thresholding is a simple yet effective feature selection method that helps exclude low-variance features, reduce noise, and optimize input data.

The aim of this study was to develop a computer-aided diagnosis (CAD) method to assist clinicians in diagnosing lumbosacral pathologies. Plain lumbosacral radiography is primarily used for low back pain. Plain radiography is an easily accessible, inexpensive, and low-radiation method. Physicians may not have sufficient experience to evaluate plain lumbosacral radiographs, and some findings may be overlooked for reasons such as workload or carelessness. Clinicians can receive objective assistance in evaluating plain lumbosacral radiographs from the proposed CAD model.

MATERIAL AND METHODS

An overview of the research architecture is presented in Figure 1.

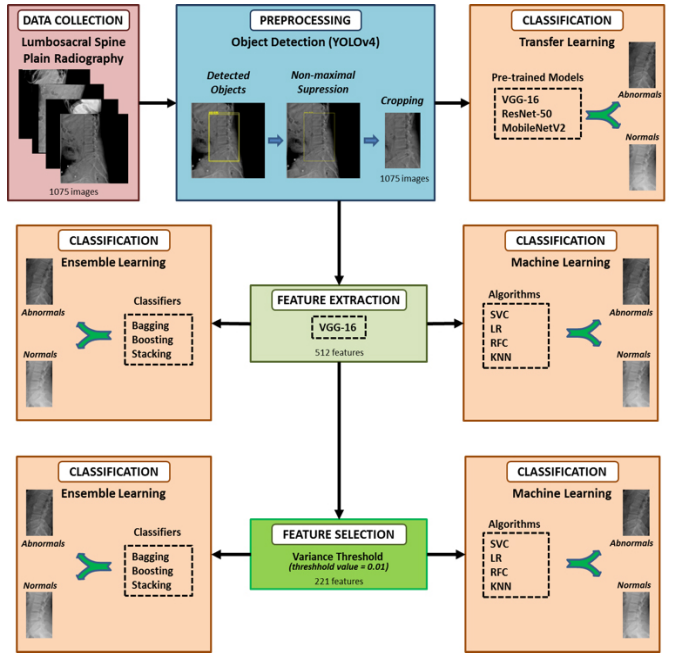


Figure 1. Overview of proposed CNN research architecture
CNN: Convolutional neural network

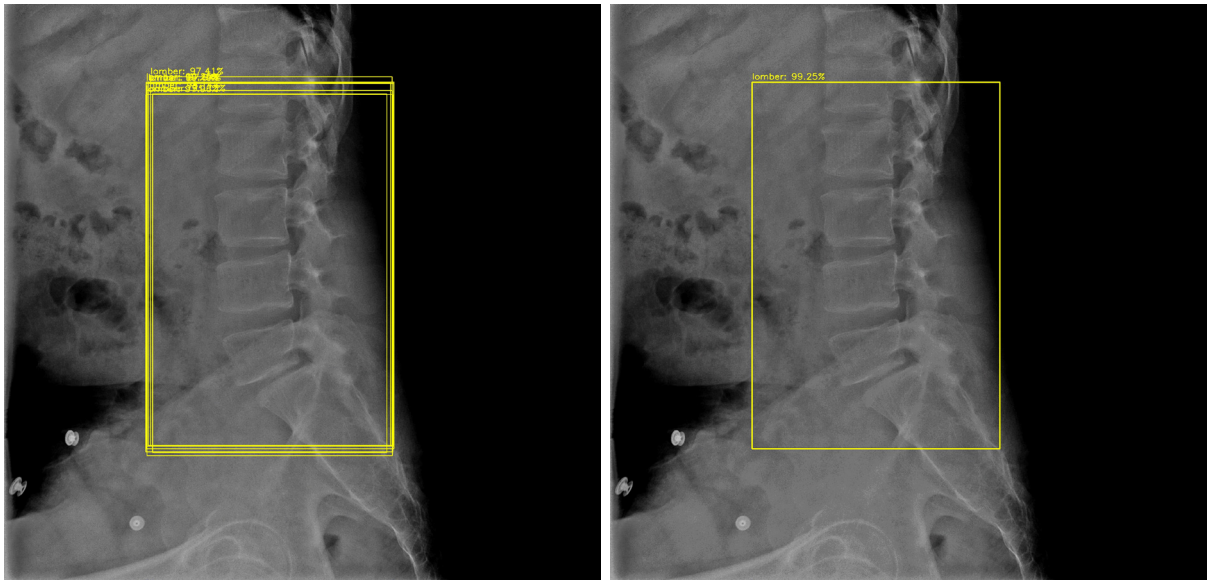
Dataset

A dataset of images was obtained from plain lumbosacral radiographs of patients examined between January 1, 2020, and March 1, 2022 at the rheumatology outpatient clinic of University of Health Sciences Turkey, Ankara Bilkent City Hospital. Approval with a waiver of informed consent, including written permission from the radiology department, was obtained from the University of Health Sciences Turkey, Ankara Bilkent City Hospital Clinical Research Ethics Committee for the study (date: 09/03/2022, number: E1-22-2546). Radiographs with low image quality (a total of 122) were excluded before obtaining the final dataset, which contained 633 abnormal and 442 normal lateral lumbosacral radiographs. The lumbosacral radiographs were labeled independently by one rheumatology specialist and one radiology specialist with more than 10 years of experience. These authors did not include radiographs that were not labeled as belonging to the same class in the study.

Data Pre-processing

The radiographic images used in this study varied in dimensions, ranging from 300x2020 to 800x2020 pixels. In plain lumbosacral radiographs, various artifacts, such as patient name, date, number, and direction, could adversely affect training. The first lumbar vertebra and sacrum had to be cropped from the entire image to discard noisy areas that were unnecessary for training and to shorten the training period. The dataset contained 1075 images, and manually cropping these images would have been a labor-intensive and time-consuming task. The YOLO algorithm was used to automate the cropping process. A total of 30 images (24 for training and six for validation) were labeled in YOLO format by the rheumatologist. In addition, the YOLOv4 configuration file was adjusted to accommodate a single class. The YOLOv4 training parameters were configured with the following settings: batch size of 16, 8 subdivisions, momentum of 0.9, and learning rate of 0.001. The pre-trained Darknet53 YOLOv4 weights were utilized to retrain the network. After 2,000 iterations, the Keras TensorFlow environment was used to create an object detector with the obtained weights. A non-maximal suppression algorithm was employed to crop the bounding box regions automatically from all dataset images. The database stored the edited images, which were then classified. Two images of a patient's radiograph are displayed in Figure 2: (a) bounding boxes and (b) the final clipping rectangle determined after applying the non-maximum suppression algorithm. The image dataset was divided randomly into training (70%), validation (15%), and test sets (15%). The image distribution is presented in Table 1.

The hyperparameter tuning process utilized the validation dataset, and the model accuracy was assessed using the test set. The stochastic gradient descent with momentum technique was used for the optimization process. The other hyperparameters



(a)

Bounding boxes obtained by applying YOLOv4 to the image

(b)

The cropping rectangle calculated after the non-max suppression algorithm

Figure 2. Outputs of feature selection process
YOLO: You Only Look Once

for the experiments are set as follows: epoch is 25, the validation frequency is 16, mini batch size =16, the L2 regularization is set to 0.004, and initial learning rates =0.0003.

In our study, the variance thresholding method was applied to enhance the interpretability and efficiency of our CAD model, ensuring that selected features were meaningful for accurate diagnosis of lumbosacral pathology.

Data Processing Environment

The research was conducted using a computer with an Intel® Core™ i7-9750H CPU @ 2.60 GHz processor, an NVIDIA GeForce RTX 2060 graphics card, and 32 GB RAM, and it was operated on a 64-bit Windows 10 system. Python 3.9 was the programing language utilized within the Keras TensorFlow environment. Essential libraries were imported, and statistical calculations were performed using the Scikit-learn library.

Transfer Learning, Data Augmentation

Pre-trained models with varying characteristics, which were trained using natural images from the ImageNet dataset, were suitable for transfer learning. This study used pre-trained VGG-16, ResNet-50, and MobileNetV2 models for transfer learning.

Data augmentation can be applied when insufficient data are available. For the purpose of augmentation, we added slightly modified versions of the existing data to the dataset to enhance the model's accuracy and avoid overfitting. Rotation, translation, and flipping transformations were used for the images in this study for augmentation.

Statistical Analysis

Each model's performance was assessed using metrics such as accuracy, sensitivity, specificity, precision, F1 score, area under the receiver operating characteristic (ROC) curve (AUC), and Cohen's kappa coefficient. The confusion matrix and ROC curve were used to calculate these metrics. The deep learning toolbox was used to test the models and obtain a confusion matrix (TP: True positive; FP: False positive; TN: True negative; FN: False negative).

RESULTS

The transfer learning method utilized pre-trained VGG-16, ResNet-50, and MobileNetV2 models. The models were tested on the test dataset after training. The accuracy, sensitivity, specificity, precision, F1 score, AUC, and Cohen's kappa coefficient values obtained for the VGG-16, ResNet-50, and MobileNetV2 models are presented in Table 2. The confusion matrix and ROC curve resulting from testing the VGG-16 model are depicted in Figure 3. Figure 4 shows the prediction results for four randomly selected images during testing with the VGG-16 model. A technique called gradient-weighted class activation mapping (Grad-CAM) was used to generate heatmaps that highlight important decision-making regions in the model (25,26). Figure 5 shows a lumbosacral plain radiography image obtained with Grad-CAM, indicating the regions that were important for model prediction.

Feature extraction was performed from the intermediate layer of VGG-16, resulting in 512 features. A variance threshold was then applied, resulting in 404 selected features with a variance threshold of zero and 221 selected features with a variance threshold of 0.01. Subsequently, ensemble learning models (Bagging, Boosting, and Stacking), and machine learning models [support vector classifier, logistic regression (LR), random forest classifier, and k-nearest neighbours (KNN)] were trained using both sets of all 512 extracted features and these 221 selected features separately. The ensemble model hyperparameters are shown in Table 3, the performance metrics are given in Tables 4-6 are for machine learning. Figure 6 shows the performance scores of the ensemble learning models before and after feature selection, and Figure 7 shows the performance scores of the machine learning. The suffix "b" denotes the scores prior to feature selection, and "a" represents the scores after feature selection.

Table 1. Numbers of images used for training, validation, and testing

	Training	Validation	Test	Total
Abnormal	458	80	95	633
Normal	319	56	67	442

Table 2. Performance metrics of VGG-16, ResNet-50, and MobileNetV2 pretrained models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)	AUC	Kappa
VGG-16	89.5	92.5	87.3	83.7	87.9	0.95	0.78
ResNet-50	84.5	88.0	82.1	77.6	82.5	0.92	0.68
MobileNetV2	80.8	85.0	77.8	73.0	78.6	0.84	0.61

AUC: Area under the curve

The overall performance of each model can be interpreted using the radar chart shown in Figure 8. Considering the assumptions "shape's size can reveal the model's power" and "the bigger the shape, the higher the performance", the ensemble learning method using the stacking classifier utilizing only the selected features can be the most robust model.

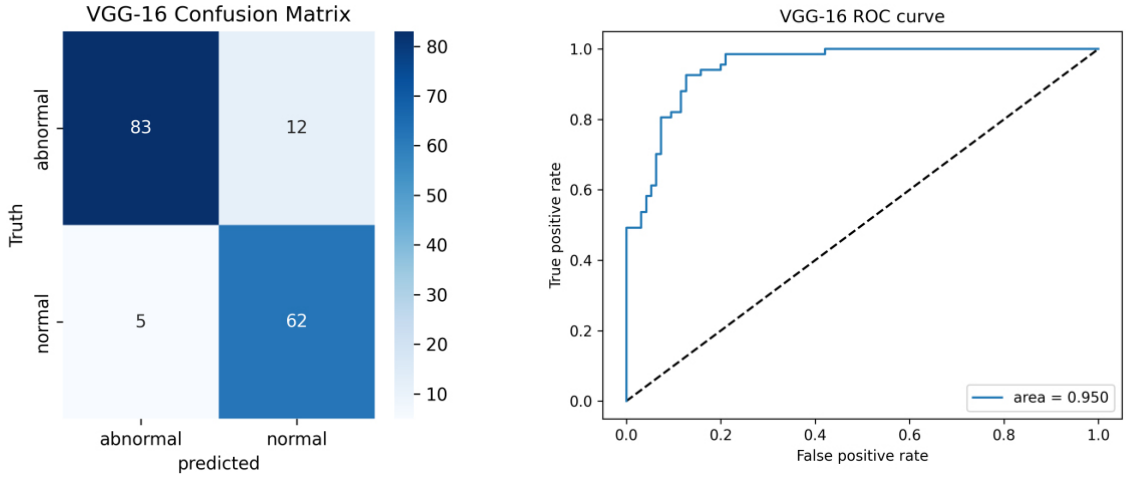


Figure 3. Confusion matrix (left) and receiver operating characteristic curve (right) obtained during testing of the VGG-16 model
ROC: Receiver operating characteristic

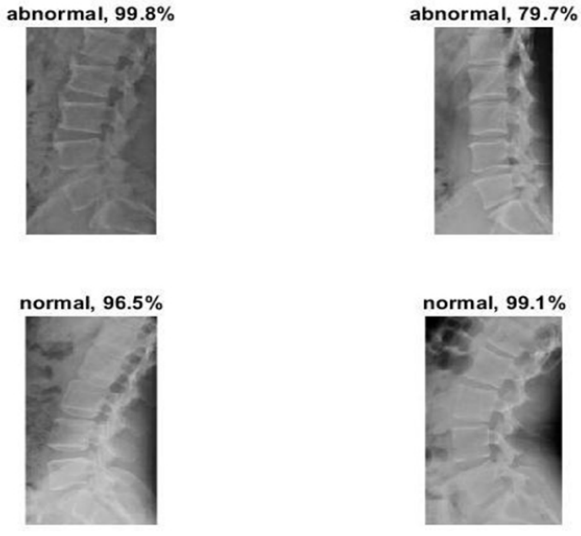


Figure 4. Prediction results from six randomly selected images during testing of the VGG-16 Model

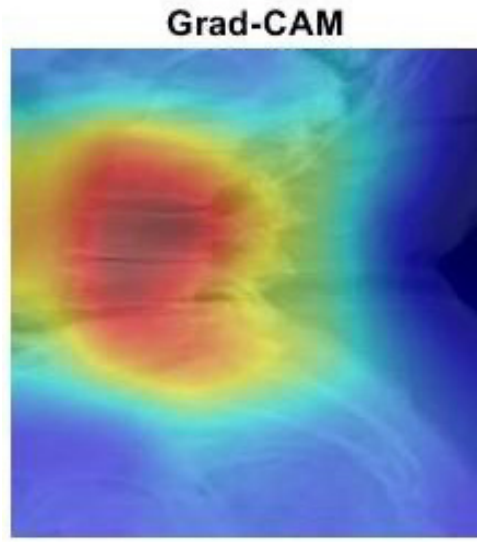


Figure 5. Lateral lumbar radiography image obtained using the Grad-CAM technique

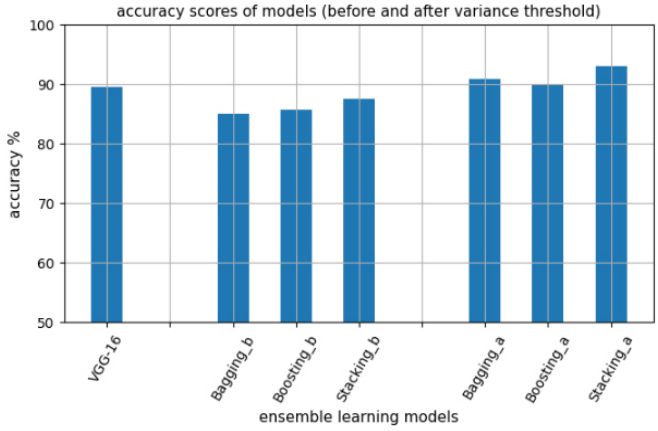
Table 3. Hyperparameters used in ensemble learning

Classifier	Parameters
Bagging	*BaggingClassifier(base_estimator = random_forest, n_estimators=100, random_state= 10)* *RandomForestClassifier(min_samples_leaf= 1, n_estimators=500, max_features = 2, max_depth = 100, bootstrap =True)*
Boosting	*AdaBoostClassifier(DecisionTreeClassifier(max_depth=1), n_estimators=200*
Stacking	*estimators = [('rf', RandomForestClassifier(n_estimators=10, random_state= 2)),* *('knn',KNeighborsClassifier(n_neighbors=5))], Meta_estimator = logistic regression*

DISCUSSION

Plain lateral lumbosacral radiographs were used in this study to diagnose lumbar pathologies, such as disc pathologies, spondylolisthesis, and osteoarthritic changes. The YOLOv4 object detector algorithm was used to eliminate artifacts not required for training the radiographs. The object detector automatically cropped all radiographs to isolate the lumbar and sacral vertebrae, which are the regions of interest. The transfer learning application involved the use of pretrained VGG-16, ResNet-50, and MobileNetV2 networks for object classification. The evaluation of each model's performance was based on metrics such as accuracy, sensitivity, specificity, precision, F1 score, AUC, and Cohen's kappa coefficient.

In distinguishing lumbar vertebral pathologies, identification of abnormal case radiographs using features selected using a threshold value from features extracted by processing the intermediate layer of VGG-16 outperformed transfer learning



"b" stands for before feature selection, "a" stands for after feature selection

Figure 6. Performance scores of the ensemble learning models

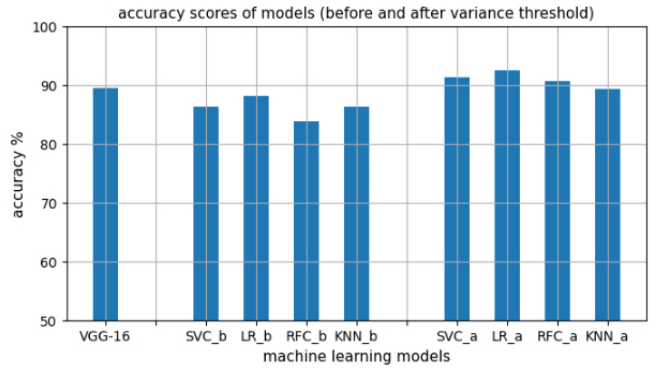


Figure 7. Performance scores of the machine learning models SVC: Support vector classifier, LR: Logistic regression, RFC: Random forest classifier, KNN: k-nearest neigbors, b: Before feature selection, a: After feature selection

Table 4. Performance metrics of ensemble learning algorithms

Feature selection	Classifier	Accuracy (%)	Sensitivity (%)	Specificity(%)	Precision (%)	F1 score (%)	AUC	Cohen's kappa
Before	Bagging	85.0	77.6	90.4	85.2	81.2	0.94	0.68
	Boosting	85.7	85.0	86.1	81.1	83.2	0.93	0.70
	Stacking	87.5	85.0	89.3	85.0	85.0	0.93	0.74
After*	Bagging	90.9	88.3	93.1	91.5	89.9	0.97	0.81
	Boosting	89.8	91.8	88.2	86.8	89.2	0.97	0.79
	Stacking	93.0	91.8	94.1	92.9	92.3	0.97	0.86

*After feature selection with variance threshold: 0.01, AUC: Area under the curve

Table 5. Hyperparameters used in machine learning

Classifier	Parameters
RF	*RandomForestClassifier(n_jobs=-1, class_weight='balanced', max_depth= 5, random_state=41)*
SVM	*SVC(probability=True), default parameters*
LR	*LogisticRegression(solver='lbfgs', max_iter=500, random_state=12)*
KNN	*KNeighborsClassifier(n_neighbors= 4)*

RF: Random forest, SVM: Support vector machine, LR: Logistic regression, KNN: k-nearest neighbors, SVC: Support vector classifier

Table 6. Machine learning algorithm performance metrics before and after feature selection

Feature selection	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 score (%)	AUC	Cohen's kappa
Before	SVC	86.3	82.0	89.3	84.6	83.3	0.94	0.71
	LR	88.1	85.0	90.4	86.3	85.7	0.95	0.75
	RFC	83.8	82.0	85.1	79.7	80.8	0.93	0.66
	KNN	86.3	76.1	93.6	89.4	82.2	0.93	0.71
After*	SVC	91.2	90.4	91.9	90.4	90.4	0.96	0.82
	LR	92.5	93.1	91.9	90.6	91.8	**0.97**	0.84
	RFC	90.6	89.0	91.9	90.2	89.6	0.96	0.81
	KNN	89.3	82.1	**95.4**	**93.7**	87.5	0.96	0.78

*After feature selection with variance threshold: 0.01, SVC: Support vector classifier, LR: Logistic regression, RFC: Random forest classifier, KNN: k-nearest neighbors

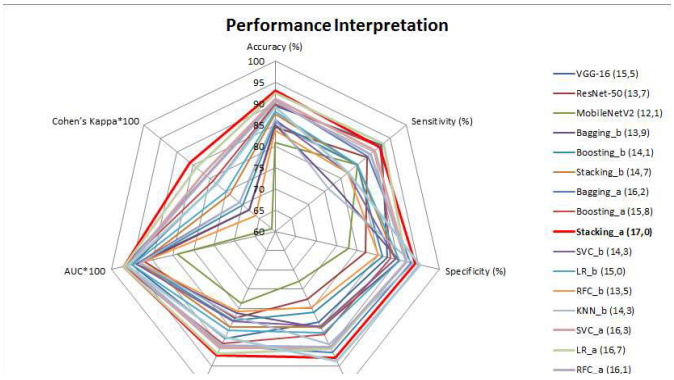


Figure 8. Performances of the models. Area of each shape is shown in parenthesis
SVC: Support vector classifier, LR: Logistic regression, RFC: Random forest classifier, KNN: k-nearest neigbors, b: Before feature selection, a: After feature selection

models using pre-trained networks, such as VGG-16, ResNet-50, and MobileNetV2, and learning methods that do not apply feature selection. The results from the ensemble learning model with the stacking classifier using features selected using a threshold value of 0.01 from the extracted features were as follows: accuracy, 93.0% (best); sensitivity, 91.8%; specificity, 94.1%; precision, 92.9%; F1 score, 92.3% (best); AUC, 0.97 (one of the best); and Cohen's kappa, 0.86 (best). The results from machine learning model with KNN classifier using the same selected features set were: accuracy 89.3%, sensitivity 82.1%, specificity 95.4% (best), precision 93.7% (best), F1 score 87.5%, AUC 0.96, and Cohen's kappa 0.78 and the results from machine learning model with LR classifier using the same set were: accuracy 92.5%, sensitivity 93.1% (best), specificity 91.9%, precision 90.6%, F1 score 91.8%, AUC 0.97 (one of the best), and Cohen's kappa 0.78.

Many successful studies have been conducted on plain radiographs using deep learning methods. Üreten et al. (27) studied normal (n=290) and sacroiliitis pelvic radiographs (n=295), in which pre-trained VGG-16 ResNet-101 and Inception-101 architectures were used for the deep learning models. The test images yielded 89.9%, 90.9%, 88.9%, 88.9%, and 0.96 for the accuracy, sensitivity, specificity, precision, and AUC performance measures, respectively (27). Another study by Üreten et al. (9) evaluated hip osteoarthritis using a transfer learning application with the VGG-16 network, using 221 normal hip X-rays and 213 osteoarthritis hip X-rays. Values of 90.2%, 97.6%, 83.0%, and 84.7% were obtained for the accuracy, sensitivity, specificity, and 84.7%, respectively (9).

Cina et al. (28) achieved success with absolute median errors of 1.84°, 2.43°, and 1.98° for the L1-L5, L1-S1, and SS angles, respectively, using a deep learning model for the localization of thoracolumbar vertebrae using 10,193 images. Another study based on a deep learning model using a total of 871 images, consisting of 413 X-ray and 458 magnetic resonance imaging (MRI), in which lumbar vertebral imaging findings were evaluated using MRI and X-ray in patients with low back pain, values of 97% for specificity, 94% for sensitivity, and 0.98 for AUC performance were obtained (29). Studies have also been conducted to determine the lumbar lordosis angle (30,31) and lumbar spondylolisthesis using plain radiographs with the deep learning method (32,33). Deep learning has also been applied to detect vertebral compression fractures (34,35).

One of the review articles on deep-learning studies using lumbar, cervical, and thoracic vertebral images conducted between 2006 and 2020 stated that deep-learning methods have enormous potential and can help clinical staff improve the level of medical care, increase work efficiency, and reduce the incidence of adverse events (36).

A deep learning approach and the VGG-16 architecture were used to analyze 161 normal and 170 lateral cervical radiographs of osteoarthritis and degenerative disc disease in a previous study. The study has an accuracy of 93.9%, sensitivity of 95.8%, specificity of 92.0%, and precision of 92.0%. In that study, pre-processing was performed manually, and regions that were not necessary for training due to the noise they contained were clipped from each radiograph (37). Deep learning-based object detection utilizes the R-CNN family, single-shot detector, and YOLO algorithms (11,21,38). In this study, the YOLOv4 algorithm was trained on 30 lumbar radiographs. With this model, radiographs were automatically cropped, and regions that were not required for training and that could adversely affect the results were removed. Thus, an end-to-end model was obtained, and classification was then performed on the radiographs.

In the present study, we applied transfer learning methods using pre-trained VGG-16, ResNet-50, and MobileNetV2 networks. Although transfer learning methods allow training on fewer data (39,40), overfitting is a fundamental problem. In our study, dropout, learning rate decay, L2 regularization, and early stopping were applied to prevent overfitting, and we did not observe overfitting in the training-test graphics and results. When deep learning methods are used, it is not known which features the method learns (black box). Hence, heat maps can be created using the GradCAM method to determine which region of the image the deep learning algorithm is recognizing; this technique highlights regions that are important in the decisions made by the model (25,26). In this study, Grad-CAM techniques were used to create heatmaps.

The use of imaging methods has become more frequent in recent years because radiologists cannot evaluate plain radiographs in most centers because of workload pressures. Machine learning and deep learning models offer ways to help clinicians in this regard. To develop models suitable for clinical use, multicenter studies using a large number of radiographs are needed. The limitations of this study include the small number of radiographs and the fact that classification was conducted using radiographs obtained from a single center.

Many pathologies are related to low back pain, heavy lifting, muscle and ligament tension due to sudden movements, degenerative disc pathologies, osteoarthritic changes, skeletal disorders such as scoliosis and spondylolisthesis, as well as fibromyalgia. The proposed method is helpful for diagnosing pathologies that can be detected by plain radiography. However, it cannot be used to diagnose fibromyalgia, which is an essential cause of low back pain. The diagnosis of fibromyalgia is based on the patient's information as well as the presence of trigger points (41).

The performance of the deep learning methods improved as the number of images increased. If studies were conducted on images obtained from different centers and with different X-ray devices, it would be possible to generalize the developed model.

CONCLUSION

This study investigated the possibilities of improving the performance of machine learning methods via feature selection with variance thresholding. The proposed model appears promising because it can assist clinicians in evaluating plain radiographs, which is a promising first step in the management of lower back pain.

Footnote

Ethics Committee Approval: The study was obtained from the University of Health Sciences Turkey, Ankara Bilkent City Hospital Clinical Research Ethics Committee for the study (date: 09/03/2022, number: E1-22-2546).

Informed Consent: Since this study included retrospective research on archive radiographs and did not publish the patients' personal information, obtaining consent forms was unnecessary.

Authorship Contributions

Concept: Y.M., S.D., K.O., E.A., K.Ü., H.H.M., Design: Y.M., A.K., S.D., K.O., E.A., E.S.S., K.Ü., H.H.M., Data Collection or Processing: Y.M., A.K., S.D., E.S.S., K.Ü., H.H.M., Analysis or Interpretation: Y.M., K.Ü., H.H.M., Literature Search: Y.M., A.K., S.D., K.O., E.A., E.S.S., Writing: Y.M., A.K., E.S.S., K.Ü., H.H.M.

Conflict of Interest: The authors have no conflicts of interest to declare.

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

REFERENCES

1. Hoy D, Brooks P, Blyth F, et al. The epidemiology of low back pain. Best Pract Res Clin Rheumatol. 2010;24:769-81.
2. Kerry S, Hilton S, Dundas D, et al. Radiography for low back pain: a randomised controlled trial and observational study in primary care. Br J Gen Pract. 2002;52:469-74.
3. Raastad J, Reiman M, Coeytaux R, et al. The association between lumbar spine radiographic features and low back pain: a systematic review and meta-analysis. Semin Arthritis Rheum. 2015;44:571-85.
4. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci USA. 2018;115: 11591-6.

5. Lim J, Kim J, Cheon S. A deep neural network-based method for early detection of osteoarthritis using statistical data. Int J Environ Res Public Health. 2019;16:1281.
6. Cheng CT, Ho TY, Lee T, 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. Ho-Le TP, Center JR, Eisman JA, et al. Prediction of hip fracture in post-menopausal women using artificial neural network approach. Annu Int Conf IEEE Eng Med Biol Soc. 2017;2017:4207-10.
8. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med. 2019;2:31.
9. Üreten K, Arslan T, Gültekin KE, et al. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiol. 2020;49:1369-74.
10. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE Computer Society. Las Vegas, NV, USA: 2016.
11. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. 2020.
12. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
13. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. Institute of Electrical and Electronics Engineers Transactions on Medical Imaging. 2016;35:1153-9.
14. Anwar SM, Majid M, Qayyum A, et al. Medical image analysis using convolutional neural networks: a review. J Med Syst. 2018;42:226.
15. Üreten K, Maraş HH. Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods. J Digit Imaging. 2022;35:193-9.
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the Association for Computing Machinery. 2017;60:84-90.
17. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: prooceding of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE; 2015.
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014.
19. Howard AG, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. 2017.
20. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: prooceding at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI, USA: 2017.
21. Aly GH, Marey M, El-Sayed SA, et al. YOLO based breast masses detection and classification in full-field digital mammograms. Comput Methods Programs Biomed. 2021;200:105823.
22. Singh S, Ahuja U, Kumar M, et al. Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. Multimed Tools Appl. 2021;80:19753-68.
23. Nie Y, Sommella P, O'Nils M, et al. Automatic detection of melanoma with Yolo deep convolutional neural networks. In: proceeding at the 2019 E-Health and Bioengineering Conference (EHB); Iasi, Romania: 2019.
24. Raihan-Al-Masud M, Mondal MRH. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. PLoS One. 2020;15:e0228422.
25. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: proceeding at the 2017 IEEE International Conference on Computer Vision Venice, Italy: (ICCV); 2016.
26. Chattopadhay A, Sarkar A, Howlader P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: proceeding at the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); Lake Tahoe. NV. USA: 2018.
27. Üreten K, Maraş Y, Duran S, et al. Deep learning methods in the diagnosis of sacroiliitis from plain pelvic radiographs. Modern Rheumatol. 2021;33:202-6.
28. Cina A, Bassani T, Panico M, et al. 2-step deep learning model for landmarks localization in spine radiographs. Sci Rep. 2021;11:9482.
29. Tan WK, Hassanpour S, Heagerty PJ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. Acad Radiol. 2018;25:1422-32.
30. Cho BH, Kaji D, Cheung ZB, et al. Automated measurement of lumbar lordosis on radiographs using machine learning and computer vision. Global Spine J. 2020;10:611-8.
31. Kitamura G. Hanging protocol optimization of lumbar spine radiographs with machine learning. Skeletal Radiology. 2021;50:1809-19.
32. Nguyen TP, Chae DS, Park SJ, et al. Deep learning system for Meyerding classification and segmental motion measurement in diagnosis of lumbar spondylolisthesis. Biomed Signal Process Control. 2021;65:102371.
33. Saravagi D, Agrawal S, Saravagi M, et al. Diagnosis of lumbar spondylolisthesis using optimized pretrained CNN models. Comput Intell Neurosci. 2022;2022:7459260.
34. Seo JW, Lim SH, Jeong JG, et al. A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. Sci Rep. 2021;1113732.
35. Kim DH, Jeong JG, Kim YJ, et al. Automated vertebral segmentation and measurement of vertebral compression ratio based on deep learning in X-ray images. J Digit Imaging. 2021;34:853-61.
36. Ren G, Yu K, Xie Z, et al. Current applications of machine learning in spine: From clinical view. Global Spine J. 2020;12:1827-40.
37. Maraş Y, Tokdemir G, Üreten K, et al. Diagnosis of osteoarthritic changes, loss of cervical lordosis, and disc space narrowing on cervical radiographs with deep learning methods. Jt Dis Relat Surg. 2022;33:93-101.
38. Cheng R. A survey: Comparison between Convolutional Neural Network And YOLO in image identification. J Phys Conf Ser. 2020:1453.
39. Gong Y, Luo J, Shao H, et al. A transfer learning object detection model for defects detection in X-ray images of spacecraft composite structures. Compos Struct. 2022;284:115136.
40. Tian Y, Wang J, Yang W, et al. Deep multi-instance transfer learning for pneumothorax classification in chest X-ray images. Med Phys. 2022;49:231-43.
41. Staud R. Evidence for shared pain mechanisms in osteoarthritis, low back pain, and fibromyalgia. Curr Rheumatol Rep. 2011;13:513-20.