Published in final edited form as:

Acad Radiol. 2022 December ; 29(12): 1819–1832. doi:10.1016/j.acra.2022.02.020.

Deep Learning Classification of Spinal Osteoporotic Compression Fractures on Radiographs using an Adaptation of the Genant Semiquantitative Criteria

Corresponding Author: Nathan M. Cross MD MS, Department of Radiology, University of Washington, 1959 NE Pacific Street, Box 357115, Seattle, WA 98195-7115, USA, nmcross@uw.edu, Phone: 206 598-2870.

Authors' Contributions

NC, DH, and JGJ conceptualized and designed the study. NEL, LYL, and LM facilitated access to the MrOS dataset and advised the team throughout model development. QD did the coding implementation and evaluation of the classification model. QD and NC performed literature review and wrote the initial draft of the paper. GL and NC extensively edited and revised the paper and contributed equally to the paper. DH, NEL, LYL, LM, JGJ, and JP revised the paper and provided guidance on the project. PC, DMK, SKJ, and the Osteoporotic Fractures in Men (MrOS) Study Publications Committee revised the paper.

Conflicts of Interest:

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Nathan M. Cross reports financial support was provided by General Electric-Association of University Radiologists Radiology Research Academic Fellowship.

Qifei Dong reports financial support was provided by National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Gang Luo reports financial support was provided by National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Li-Yung Lui reports financial support was provided by National Institute of Health.

Deborah M. Kado reports financial support was provided by National Institute on Aging.

Peggy M. Cawthon reports financial support was provided by National Institutes of Health.

David Haynor reports financial support was provided by National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Jeffrey G. Jarvik reports financial support was provided by National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Nathan M. Cross reports financial support was provided by National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Deborah M. Kado reports a relationship with National Osteoporosis Foundation that includes: speaking and lecture fees.

Deborah M. Kado reports a relationship with American Bone Health that includes: speaking and lecture fees.

Deborah M. Kado reports a relationship with Interdisciplinary Symposium on Osteoporosis that includes: speaking and lecture fees.

Deborah M. Kado reports a relationship with Veterans Administration Health System that includes: travel reimbursement.

Deborah M. Kado reports a relationship with Stanford University School of Medicine that includes: travel reimbursement.

Deborah M. Kado reports a relationship with American Society of Bone and Mineral Research that includes: travel reimbursement.

Deborah M. Kado reports a relationship with ASBMR Task Force on Long-Term Safety and Efficacy of Vertebral Augmentation that includes: board membership.

Deborah M. Kado reports a relationship with Data Safety Monitoring Board, TOPAZ Trial that includes: board membership.

Deborah M. Kado reports a relationship with NIH NIA Aging Workshop for the American Society of Bone and Mineral Research (ASBMR) that includes: board membership.

Jeffrey G. Jarvik reports a relationship with GE-Association of University Radiologists Radiology Research Academic Fellowship that includes: travel reimbursement.

Gang Luo currently works part-time at Amazon as an Amazon Scholar.

Deborah M. Kado reports Wolters Kluwer/UpToDate: Royalties as a chapter author.

Jeffrey G. Jarvik reports Springer Publishing: Royalties as a book co-editor; and Wolters Kluwer/UpToDate: Royalties as a chapter author.

All other authors report no conflicts of interest.

Availability of Data and Material

MrOS study data are largely publicly available through their website https://mrosonline.ucsf.edu.

Code Availability

Our code for image pre-processing and augmentation was uploaded to https://github.com/UW-CLEAR-Center/Preprocessing_for_Spinal_OCF_Detection. Publicly available libraries, open source code, and pretrained models (e.g., Tensorflow, TF-Slim, and TensorFlow Model Garden) were used to build and train the deep learning model in this research.

Ethics Approval

The MrOS study and its associated data that have been extensively published were previously approved by the IRBs of each local site including: Birmingham AL, Minneapolis MN, Palo Alto CA, Pittsburgh PA, Portland OR, and San Diego CA. The study described in this paper was a retrospective study of this existing anonymized research dataset.

Consent to Participate

Not applicable

Consent for Publication

Not applicable

Qifei Dong, MS,
Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, USA

Gang Luo, PhD,
Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, USA

Nancy E. Lane, MD,
Department of Medicine, University of California - Davis, Sacramento, CA 95817, USA

Li-Yung Lui, MA MS,
Research Institute, California Pacific Medical Center, San Francisco, CA 94143, USA

Lynn M. Marshall, ScD,
Epidemiology Programs, Oregon Health and Science University-Portland State University School of Public Health, Portland, OR 97239, USA

Deborah M. Kado, MD, MS,
Department of Medicine, Stanford University, Stanford, CA 94305, USA; Geriatric Research Education and Clinical Center (GRECC), Veterans Administration Health System, Palo Alto, CA 94304, USA

Peggy Cawthon, PhD MPH,
California Pacific Medical Center Research Institute, Department of Epidemiology and Biostatistics, University of California - San Francisco, San Francisco, CA 94143, USA

Jessica Perry, MS,
Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Sandra K Johnston, PhD, RN,
Department of Radiology, University of Washington, Seattle, WA 98195-7115, USA

David Haynor, MD PhD,
Department of Radiology, University of Washington, Seattle, WA 98195-7115, USA

Jeffrey G. Jarvik, MD MPH,
Departments of Radiology and Neurological Surgery, University of Washington, Seattle, WA 98104-2499, USA

Nathan M. Cross, MD MS
Department of Radiology, University of Washington, 1959 NE Pacific Street, Box 357115, Seattle, WA 98195-7115, USA

Abstract

Rationale and Objectives: Osteoporosis affects 9% of individuals over 50 in the United States and 200 million women globally. Spinal osteoporotic compression fractures (OCFs), an osteoporosis biomarker, are often incidental and under-reported. Accurate automated opportunistic OCF screening can increase the diagnosis rate and ensure adequate treatment. We aimed to develop a deep learning classifier for OCFs, a critical component of our future automated opportunistic screening tool.

Materials and Methods: The dataset from the Osteoporotic Fractures in Men Study comprised 4,461 subjects and 15,524 spine radiographs. This dataset was split by subject: 76.5% training, 8.5% validation, and 15% testing. From the radiographs, 100,409 vertebral bodies were extracted, each assigned one of two labels adapted from the Genant semiquantitative system: moderate to severe fracture vs. normal/trace/mild fracture. GoogLeNet, a deep learning model, was trained to classify the vertebral bodies. The classification threshold on the predicted probability of OCF outputted by GoogLeNet was set to prioritize the positive predictive value (PPV) while balancing it with the sensitivity. Vertebral bodies with the top 0.75% predicted probabilities were classified as moderate to severe fracture.

Results: Our model yielded a sensitivity of 59.8%, a PPV of 91.2%, and an F_1 score of 0.72. The areas under the receiver operating characteristic curve (AUC-ROC) and the precision-recall curve were 0.99 and 0.82, respectively.

Conclusion: Our model classified vertebral bodies with an AUC-ROC of 0.99, providing a critical component for our future automated opportunistic screening tool. This could lead to earlier detection and treatment of OCFs.

Keywords

Osteoporosis; fragility fracture; deep learning; semiquantitative; opportunistic screening

Introduction

Osteoporosis is a debilitating disease affecting 9% of individuals over 50 years old in the United States [1] and 200 million women globally [2]. In a lifetime, one in three individuals in a developed country will incur an osteoporotic compression fracture (OCF) [2]. After the first fracture, the risk for subsequent fractures is dramatically increased [3–5]. Even a single OCF is associated with a decreased quality of life and a higher mortality rate [6].

Osteoporosis screening is underutilized despite being endorsed by many organizations, including the US Preventive Services Task Force. Between 2004 and 2006, less than 1/3 of women who should have been screened underwent bone mineral density testing [7]. Jain *et al.* [8] reported that the rate of osteoporosis screening for high-risk men is low. The cause of underutilization of osteoporosis screening is multifactorial, however Medicare payment cuts for dual-energy X-ray absorptiometry screening were associated with a screening rate decrease of 56% between 2006 and 2010 [9].

Opportunistic screening can complement current screening methods. Several groups have published approaches to opportunistic screening using pre-existing imaging to increase osteoporosis detection rates [10–26]. Many groups used computerized tomography images [10–22], while few used radiographs [23–26]. Since radiography is a ubiquitous imaging modality used early in diagnostic workup of many conditions with an estimated 183 million exams in US hospitals in 2010 [27], accurate opportunistic osteoporosis screening using radiographs is as important as that using computerized tomography. Among the groups that used radiographs, Lee *et al.* [23] and Zhang *et al.* [24] used machine learning algorithms to estimate bone mineral density. However, using bone mineral density as a biomarker

of osteoporosis detection has known limitations [28, 29]. Spinal OCFs are an additional osteoporosis biomarker. Spinal OCFs are often incidental on chest or abdominal images and frequently under-reported, resulting in under-diagnosis and under-treatment [30]. Applying automated opportunistic OCF screening to existing imaging studies could result in earlier and more extensive osteoporosis identification and treatment. Murata *et al.* [25] and Chou *et al.* [26] recently reported approaches to automatically detecting OCFs on radiographs. However, their studies had limitations, including small sample size, probable selection bias resulting in unrealistically high prevalence of the OCFs, and single center data leading to possible overfitting.

We ultimately attempt to construct an automated opportunistic screening tool to detect OCFs on radiographs. It would include at least three components: 1) image segmentation that automatically finds the vertebral bodies on a spine radiograph; 2) an image classifier that determines whether each vertebral body is fractured; and 3) a patient-level classifier that integrates the fracture status of all vertebral bodies in the spine radiograph. In this manuscript, we focus on describing the second component, the OCF classifier. For image classification, deep learning significantly outperforms other machine learning algorithms [31]. A deep learning model is a multi-layer neural network, which can extract features from unstructured data such as medical images. GoogLeNet [32] is a well-known deep learning model that we used to build our OCF classifier. Compared with other more recent deep learning models, GoogLeNet contains fewer parameters that need to be learned [33], is faster to train, and is less likely to overfit given a limited number of training data instances as is the case with OCF classification in this study. Our objective was to create an OCF classifier that could achieve an AUC-ROC of at least 0.9 for classifying vertebral bodies as moderately to severely fractured vs. normal/trace/possible fracture.

Materials and Methods

MrOS Dataset

The Osteoporotic Fractures in Men (MrOS) Study radiograph dataset was previously described in Orwoll *et al.* [34]. A de-identified copy of this dataset was obtained under a data use agreement with the San Francisco Coordinating Center. The MrOS study collected data from six US academic medical centers in Birmingham AL, Minneapolis MN, Palo Alto CA, Pittsburgh PA, Portland OR, and San Diego CA, at each of which a local IRB approved the study. All participants gave written informed consent. MrOS included 5,994 males aged 65 and older from six clinical centers in the United States and collected clinical and laboratory imaging data at the initial visit (Visit 1) and the follow-up visit (Visit 2) average 4.5 years later. At Visit 1 and Visit 2, lumbar and thoracic radiographs were obtained from 5,994 and 4,423 subjects, respectively. At Visit 1, the clinical centers provided film-based radiographs. At Visit 2, four centers provided film-based radiographs and the other two centers provided direct digital radiographs.

Cawthon *et al.* [35] annotated the radiographs in the MrOS dataset with the Genant semiquantitative (SQ) criteria [36] (see Figure 1) and identified the margin of the vertebral bodies. All film-based radiographs were digitized. In total, 20,824 radiographs were in digital forms (11,982 from Visit 1 and 8,842 from Visit 2). The radiographs of 36 (0.60%)

subjects were excluded in the MrOS study primarily due to the presence of diffuse idiopathic skeletal hyperostosis, ankylosing spondylitis, and to a lesser degree technical obscuration of the spine (overlying structures, exposure, parallax, etc.). The remaining radiographs of 5,958 subjects from Visit 1 and 4,399 subjects from Visit 2 were annotated. To outline each vertebral body in each radiograph, four corners of this vertebral body and two midpoints of the superior and inferior endplates were pinpointed.

Dataset Labeling and Partitioning

Our dataset was constructed from the MrOS radiographs [34] and Cawthon *et al.*'s annotations [35]. Each data instance is an individual vertebral body extracted from the original radiograph using the four corner points from the MrOS annotation. Depicted in Figure 1, the SQ classes were aggregated into two classes: label 0 representing a normal vertebral body or a possible mild deformity and label 1 representing a moderate to severe deformity. Reasons to simplify the SQ criteria included:

1. Our design intent is ultimately to create an automated opportunistic screening tool to screen large populations and to prompt further osteoporosis evaluation based on the prediction of a probable osteoporotic fracture. Thus, for cases with a probable fracture, a member of the healthcare team will be alerted that a finding is present and further evaluation is needed. Otherwise, if there is no probable fracture, no notification will be provided and the standard of care will be maintained. This scenario represents a binary decision of whether to notify a provider or not, and thus supports a binary classification.
2. An automated opportunistic screening tool screening a large volume of cases must identify cases with high confidence to prevent undue burden on the healthcare enterprise. The mild deformity category may include deformities that are not definitively OCFs but rather are congenital or associated with a disease other than osteoporosis. Thus, this category was grouped with the normal class.
3. The number of vertebral bodies are insufficient in some SQ classes like "moderate crush deformity" (see Figure 1).

Data instances were partitioned into the training, validation, and test sets by subject. The radiographs from one subject only appear in one of the three sets. We randomly selected 76.5%, 8.5%, and 15% of the subjects to form the training, validation, and test sets, respectively (see Figure 2). The percentage used to form the validation instances was set smaller than usual (e.g., 10% or 20%) due to the class imbalance of the dataset. For effective training, the training set needed subsampling to correct the class imbalance between label 1 and label 0 (Figure 2). To preserve the class distribution of the original population and to most closely evaluate real world performance, no corrections of the imbalance in the validation and test sets were performed. A smaller validation set allowed for a larger training set, even after downsampling the majority class in the training set. In the training set, data instances of label 0 were randomly sampled at a ratio of 2.5:1 (label 0 : label 1) to better balance the two classes. The ratio of 2.5:1 was determined in an earlier experiment, in which different ratios were tried and we chose the ratio to maximize the area under the precision-recall (PR) curve (AUC-PR) on the validation set.

Image Pre-processing and Augmentation

This section presents the image pre-processing and augmentation steps. First, we wanted to extract the image patches, each containing a vertebral body, from the radiographs. These image patches were termed vertebral patches. Each input data instance for our OCF classifier was a vertebral patch. Since the automated image segmentation tool is under development, each vertebral patch in this study was extracted using the manually annotated contour of the corresponding vertebral body. Second, we attempted to control the heterogeneity among the vertebral patches to a moderate range. Excessive heterogeneity in the dataset can confound deep learning models when extracting relevant features, while too little heterogeneity could result in poor generalizability of the trained model. Figure 3 shows the general steps of image pre-processing and augmentation. Image augmentation is a ubiquitous approach to increase the trained model's performance by creating subtly modified data instances for the model to be trained on. By definition, image augmentation was applied to only the training set. Among all steps shown in Figure 3, Steps 5 and 8 are image augmentation procedures applied to only the training set, while the other steps are image pre-processing steps applied to each of the training, validation, and test sets. Our code for image pre-processing and augmentation was uploaded to https://github.com/UW-CLEAR-Center/Preprocessing_for_Spinal_OCF_Detection. In the following, we describe the details of these image pre-processing and augmentation steps.

In each vertebral patch, we controlled the variation of three features: 1) the vertebral body's position; 2) the percentage of the vertebral patch's area occupied by the vertebral body; and 3) the vertebral body's tilt angle. The aspect ratio of each vertebral body was fixed after extracting it. Initially, horizontally flipping was performed if needed, to conform to the convention that the subject faces left (see Step 1 of Figure 3). To extract a vertebral body, the four corner points in the six-point morphological annotations were used to generate two diagonals. We built two coordinate axes with the x-axis bisecting the angle between the two diagonals connecting the four corner points (see Step 2 of Figure 3). The angle between this bisector and the x-axis of the vertebral patch defined the vertebral body's tilt angle. To keep the extracted vertebral body's aspect ratio constant, we bounded the vertebral body by a square fulfilling two requirements: 1) one side of the square is perpendicular to the aforementioned bisector; and 2) the square is the smallest square with none of the four corner points lying outside of it. Requirement 1 guarantees that the vertebral body is not tilted inside the square. Requirement 2 assures that the vertebral body is at the center of the square. This smallest square cannot always bound the whole vertebral body because of osteophytes and parallax. To avoid accidental cropping of the vertebral body and to provide surrounding image context during the extraction step, we expanded this smallest square around its center to enlarge its area by four times. If the enlarged square exceeded the boundary of the spine image, we zero padded the excess area. This enlarged square served as the vertebral patch. In summary, the extraction process assures that the vertebral body is positioned at the center of a patch, occupies one quarter of the patch, and is not tilted.

Affine transformation, a data augmentation method, was conducted during vertebral body extraction. For each vertebral patch, rotation, scaling, and translation were applied to the vertebral body sequentially. The requirements of the affine transformation are: 1) scale the

vertebral body's area by s%, where s was randomly selected from the range [81, 121]; 2) rotate the vertebral body by a degree randomly selected from the range [−5°, 5°]; and 3) translate the vertebral body along the x-axis and y-axis, each by a distance equal to a value randomly and independently sampled from the range [−0.05, 0.05] × the length of vertebral patch's edge. Affine transformation of the vertebral body inside the vertebral patch is basically the same as that of the square on the spine image (Step 5 in Figure 3). To expand (or shrink) the vertebral body's area by s%, we shrank (or expanded) the square's area by (1/s%)×100%. To rotate the vertebral body by an angle, we rotated the square by the same angle in the opposite direction. To translate the vertebral body by a distance, we translated the square by the same distance in the opposite direction. For each original square (the square after Step 4 in Figure 3), we conducted affine transformations on it four times to generate four augmented vertebral patches.

While the bone of the vertebral body is brighter than the background in most vertebral patches, "inverted patches" also exist, in which the bone of the vertebral body is darker than background. Figure 3 includes an inverted vertebral patch, which is shown to be inverted back in step 7 according to the convention used in this project (bone brighter than background and air). In our experiments, mixing these two types of vertebral patches in the dataset was one major negative factor on the deep learning model's performance (see the "Error Analysis" Section of the "Results" Section). Thus, it is necessary to invert the inverted patches to make the bone of the vertebral body in each vertebral patch brighter than the background. Figure 4 shows our algorithm for detecting inverted patches. Our intuition is that the pixels of the vertebral body's endplates should have gray intensities closer to those inside the vertebral body than to those outside. Our algorithm consists of six steps:

1. Find the vertebral body's endplates. In the vertebral patches that were not generated by affine transformation, the vertebral body is non-tilted and hence the endplates are usually close to being horizontal. To detect horizontal lines, we used a Sobel operator [37] with a kernel size of 5. Then we used hysteresis thresholding [37] to check which horizontal lines were most likely to be the endplates. The low and high levels of hysteresis thresholding were set to 0.5 and 0.8, respectively.
2. For each pixel on an endplate, obtain a vertical stripe with the pixel as its midpoint. The stripe has a width of one pixel and a height of 0.05 × the vertebral patch's edge length.
3. On each vertical stripe, obtain the highest and lowest gray intensities.
4. Calculate the differences a) between the highest gray intensity and the gray intensity of the pixel on which the endplate and the vertical stripe intersect; and b) between the gray intensity of this pixel of intersection and the lowest gray intensity. The two differences were denoted by d_h and d_l, respectively. If d_l is $<d_h$, the pixel on the endplates is closer to the "dark end" of the gray intensity histogram of all the pixels on the vertical stripe. Otherwise, the pixel on the endplate is closer to the "bright end" of this histogram.
5. Obtain all of the d_l and d_h by traversing all of the pixels on the endplates.

6. Calculate and compare the means of all d_l and of all d_h to determine whether the grayscale is inverted. If the mean of all d_l is less than that of all d_h, the pixels on the endplate's clusters more on the "dark end" of the gray intensity histogram of the pixels around and on the endplates. This indicates a dark vertebral body and bright surrounding. Conversely, if the mean of all d_l is greater than or equal to that of all d_h, we determine that the grayscale is not inverted.

To further increase the accuracy of inverted patch detection, rather than relying on the decision for a single vertebral patch, we integrated the algorithm's outputs across all vertebral patches in any given spine radiograph. More specifically, majority voting was conducted on the results of all vertebral patches in a spine radiograph. The voting result of a spine radiograph was assigned to all vertebral patches from that radiograph. If the vote was a tie, inversion of the spine image was randomly assigned.

Afterwards, we conducted two additional data augmentation steps on each vertebral patch generated by affine transformation: 1) adjust the contrast and brightness using the SigmoidContrast class in the imgaug package [38]; and 2) add Gaussian noise using the AdditiveGaussianNoise class in the imgaug package [38]. When adjusting the contrast and brightness, two parameters were required: gain and cutoff. Gain was randomly sampled from the range [4, 8]. Cutoff was randomly selected from the range [0.4, 0.6]. The standard deviation of the Gaussian noise was randomly sampled from the range [0, 39].

As a result, for each vertebral body in the training set, we obtained one original vertebral patch and four augmented vertebral patches. These five patches were used for model training. The validation and test sets only contain the original vertebral patches. Before feeding into the neural network, each vertebral patch was resized to 224×224 pixels and each pixel value was normalized by subtracting the mean and then dividing by the standard deviation of the vertebral patch [31].

Model Training

GoogLeNet [32] was used to predict the presence of OCF for each vertebral body (label 1 vs. label 0). This neural network was built using Python 3.7.6, TensorFlow 2.2.0 [39], and TF-Slim 1.2.0 [40]. The original paper describing GoogLeNet [32] used an architecture with one backbone network and two auxiliary networks. The latter was not included in the GoogLeNet code provided by TF-Slim. Thus, we added these two auxiliary networks to the source code. Our code for training the GoogLeNet model was adapted from online open-source code [41].

Transfer learning [31], a commonly used technique to boost deep learning models' performance in image classification tasks, was used. In transfer learning, deep learning models pre-trained on large general-purpose imaging datasets are used as a starting point for training the model for another task, in this case classifying vertebral bodies. TensorFlow Model Garden [42] provided a pre-trained GoogLeNet model implemented in TF-Slim and pre-trained on the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset [43]. The model's output layer was adapted for binary classification to fit OCF classification. Except for the parameters for the output layer, all of the other

parameters were initialized from the pre-trained model. The parameters for the output layer and the two auxiliary networks were initialized using He initialization [44]. ILSVRC2012 uses RGB images with three channels. To use the pre-trained weights with the single grayscale channel of each vertebral patch, this channel was copied into the RGB channels.

When training GoogLeNet, we did not freeze any layer. In our experiments, freezing some layers had a limited impact on our model's performance for OCF detection. The Adam learning rate optimization [45] was used with a weighted cross entropy loss function [46], which penalizes false positives more heavily than false negatives, and the optimization was iterated over a batch size of 20. Early stopping was used to avoid overfitting [47]. After each epoch, performance was evaluated on the validation set by calculating the AUC-PR. For a highly imbalanced dataset, the AUC-PR is a more suitable performance metric than the area under the receiver operating characteristic curve (AUC-ROC) and accuracy [48]. If the AUC-PR did not increase in any of the subsequent 10 epochs, training was terminated.

GoogLeNet contains several hyper-parameters. The four hyper-parameters shown in the Appendix were tuned by random search [47] for 1,500 rounds, with the goal of maximizing the AUC-PR on the validation set. The other hyper-parameters not mentioned in the Appendix were set to their default values given in the original GoogLeNet paper [32].

Hyper-parameter tuning was done on two Ubuntu Linux servers concurrently: 1) Xeon E5-2630 with 256 GB memory and four Nvidia GeForce TITAN Xp graphics processing units (GPUs), and 2) Xeon Gold 5215 with 96 GB memory and four Nvidia GeForce 2080 Ti GPUs. The final model was trained and tested on the first server using one GPU.

Model Evaluation

We tried the following two thresholding methods to determine the classification threshold on GoogLeNet's predicted probabilities of having label 1 (moderate to severe fracture):

1. Method 1: Manually select a threshold to prioritize the positive predictive value (PPV) for the opportunistic screening use case. In the setting of a screening tool screening large volumes of studies, a tool with too many false positives could unduly burden the healthcare system. Thus, the PPV of the model should not be small. In consultation with local clinicians, a PPV of approximately 90% was targeted. To obtain a PPV of approximately 90%, the threshold was set to classify vertebral bodies with the top 0.75% predicted probabilities as label 1.
2. Method 2: Obtain the threshold by maximizing Youden's J statistic, which balances sensitivity and specificity.

Using each thresholding method, we evaluated the final model by computing seven performance measures on the test set: accuracy, sensitivity, specificity, PPV, negative predictive value (NPV), false discovery rate (FDR = 1- PPV), and F_1 score. Using the second thresholding method, these seven performance measures were also computed on the training set as well as on the validation sets. Using the test set, the ROC curve and the PR curve were plotted, the AUC-ROC and the AUC-PR were calculated. For each performance

measure, its 95% confidence interval (CI) was obtained using 2,000-fold bootstrapping analysis.

Error Analysis

During model development, an error analysis was performed to find confounders resulting in misclassification, which guided later image pre-processing. From the validation set, 150 correctly and 120 incorrectly classified vertebral patches were randomly selected. Of these, 50 were reviewed by two radiologists for visual features that might cause misclassification. These two radiologists blindly and independently reviewed all 270 vertebral patches for each confounder in each patch. The ratio of misclassification odds with and without each confounder was calculated to determine the relative significance.

Results

MrOS Dataset and Dataset Partitioning

Within the MrOS dataset, a small number of vertebral bodies were dropped because they were not annotated, usually vertebra on the edges of the film. We finally obtained 4,461 subjects with 8,915 radiographs from Visit 1 and 3,309 subjects with 6,609 radiographs from Visit 2. Table 1 shows the characteristics of the subjects in the entire, training, validation, and test sets of the MrOS dataset. Recall that the MrOS dataset is imbalanced with far more vertebral bodies with label 0 than vertebral bodies with label 1. To balance the training set, vertebral bodies with label 0 were downsampled. Both the characteristics of the subjects in the training set before downsampling and those in the training set after downsampling are shown in Table 1.

In total, 100,828 vertebral bodies were identified from all of the radiographs from both visits. Of these vertebral bodies, 376 were labeled “cannot evaluate,” “missing,” or “not applicable” were discarded. The remaining 100,452 vertebral bodies were scored using the Genant SQ criteria [36]. Of the 100,452 vertebral bodies, 43 had two annotations, which always resulted in the same label under the simplified categorization in Figure 1. The remaining 100,409 vertebral bodies were consisted of 69,453 lumbar vertebral bodies and 30,956 thoracic vertebral bodies. The numbers of thoracic and lumbar vertebral bodies in each class are listed in Figure 1.

Figure 2 shows the dataset partitioning and the final distributions of the subjects, radiographs, and vertebral bodies in the training, validation, and test sets. Figure 5 shows the number of vertebral bodies of each SQ class at each anatomic level of the spine.

Error Analysis

Table 2 shows the error analysis results for the deep learning model trained in the early stage of this study. Substantial noise was found to be the confounder with the highest odds ratio for incorrect classification. Image inversion was another prominent confounder in this sample. For each of the confounders like “Overlying metal…” and “Point placement wrong…,” the number of sampled vertebral patches with this confounder was too small to adequately assess whether this confounder can degrade the classification accuracy.

Model Evaluation

Final model training took approximately one hour. After the model was trained, a single vertebral patch inference to predict fracture or no fracture took an average of two milliseconds.

Figure 6 presents the final model's performance on the test set. Parts A and B of Figure 6 show the ROC curve with an AUC-ROC of 0.99 and the PR curve with an AUC-PR of 0.82, respectively. Part C of Figure 6 shows the model's performance measures when the classification threshold on GoogLeNet's predicted probabilities of having label 1 was determined by each of the two methods presented in the "Model Evaluation" section of the "Materials and Methods" section. Setting the classification threshold by prioritizing the PPV while balancing it with sensitivity, our model yielded a sensitivity of 59.8%, a specificity of 99.9%, a PPV of 91.2%, an NPV of 99.5%, an FDR of 8.8%, an F_1 score of 0.72, and an accuracy of 99.5%.

Table 3 shows the model's performance measures on the test set with different thresholds applied to the output probability of fracture.

Table 4 shows our model's performance measures on the validation set, the training set without the augmented vertebral patches, and the training set with the augmented vertebral patches. To compute the performance measures on each set, we set the classification threshold on the predicted probabilities of having label 1 by maximizing Youden's J statistics.

Discussion

Principle Findings

This deep learning model for detecting OCFs of individual vertebral bodies demonstrates high performance. It was developed using a large, well validated, and prospectively acquired multicenter dataset utilizing the widely used Genant SQ criteria. Final performance of AUC-ROC of 0.99 exceeds our objective of having an AUC-ROC of 0.9.

The error analysis done during model development showed that "substantial noise" was a leading cause of misclassification. Subsequently, noise was added during image augmentation to generate augmented vertebral patches. This improved the generalizability of the model to noisy patches. Another leading confounder causing misclassification was "inverted grayscale of the vertebral patch." To address this, we designed an algorithm that detects and invert the inverted vertebral patches using conventional image processing techniques (Step 7 in Figure 3). The confounders affecting few vertebral patches were disregarded.

Our ultimate goal is to create an automated opportunistic screening pipeline for large numbers of lateral clinical radiographs that image part of the spine. This deep learning model is a component of that pipeline. As mentioned in the Introduction section, multiple components are needed for the pipeline, but a model for determining the presence or absence

of fracture for each vertebral body is the most clinically critical component of the pipeline. Thus, it is critical to ensure high performance of this model.

Now that a deep learning model for classifying vertebral bodies is created, there is flexibility on how best to use the model's output, a predicted probability of osteoporotic fracture. Since this is a probability between 0 and 1 for every vertebral body processed, any value between 0 and 1 can be used to threshold the output to create the final output of whether the vertebral body has a significant fracture or not. The threshold value selected determines the sensitivity, the specificity, and the PPV of the model. The range in performance is represented in the ROC and PR curves depicted in figure 6. The PR curve exhibits a gentle slope on the left of the curve where the sensitivity is around 0.0 – 0.5 (Figure 6). This is useful, as the goal is to maintain a high precision (PPV). This region of the curve allows us to maintain a high precision with little sacrifice in recall (sensitivity) until a sensitivity of around 0.5, where the PPV starts to fall steeply. The thresholds determined by Methods 1 and 2 could be employed clinically, but ultimately the selected threshold should be driven by the use case.

Since the ultimate goal is for this deep learning model to function in a screening tool for a large number of exams where a positive result will trigger action in the healthcare system, it is important to ensure that only a minimum of false positives are generated. The exact number of false positives that would be acceptable to a healthcare system is likely to vary by healthcare system and is the subject of a separate body of work. All diagnostic tests generate false positives. The challenge is to determine how many false positives are acceptable for the use case by a specific healthcare system. Initial interviews with clinical faculty and staff suggest that a ratio of 1 false positive for every 10 true positives would not be overly burdensome. This would represent a PPV of about 91% or an FDR of about 9%. The inherent compromise is that sensitivity is sacrificed by maintaining a high PPV. These numbers were used to guide the selection of the threshold determined by Method 1 and shown in Figure 6 resulting in a sensitivity of 59.8%, a PPV of 91.2%, and an FDR of 8.8%. Our future opportunistic screening tool will supplement the current standard of care. Having a lower sensitivity could cause some osteoporotic fractures to be missed, but ultimately will not degrade the current standard of care. In other words, even if only a small number of osteoporotic fractures are detected, this tool will only improve the current standard of care since it is not replacing any stage in current clinical practice and it is using existing data acquired for other purposes.

Limitations & Future Directions

This study has several limitations:

1. The dataset came from the multicenter MrOS study that was started in 2000. Radiographs acquired by more recent digital detectors and systems could have characteristics and quality that are different from those in this dataset. We are currently creating additional datasets that include a local dataset with radiographs acquired using more modern techniques. These datasets can be used to test generalizability of our model or to train a better model.

2. The MrOS study, by design, only included male subjects from six clinical centers in the U.S. Further testing is needed to ensure that this methodology is generalizable to women and international populations. Additional datasets with female subjects and international content are currently being developed to test generalizability of our model and to train a more robust model.

3. Some studies show that the Genant SQ criteria have limitations when assessing OCFs [49]. Subtle anterior wedging has overlap with other conditions that can be mistaken for subtle OCF in the Genant SQ criteria. Future work will include using other OCF classification methods, such as the modified algorithm-based qualitative [49] approach.

4. In this study, GoogLeNet was used to build the model. Using other deep learning models such as ResNet [50] could better detect OCFs. More radiograph specific forms of augmentation could be used to model imaging chain artifacts or to approximate subject positioning variations.

5. We used spine radiographs to build the OCF classifier. Since spine radiographs are optimized to show the bones, this type of radiograph is a reasonable choice for our initial study. In the future, we will test generalizability to other exam types such as chest and abdominal radiographs.

6. Our current deep learning model is applicable only to individual vertebral bodies. To adapt to a real clinical setting using spine radiographs each having multiple vertebral bodies, we need to do additional work to create a framework for automated image segmentation and vertebral corner point identification that can feed into a vertebral patch fracture classifier.

Conclusions

Our model classified individual vertebral bodies with an AUC-ROC of 0.99, showing high performance at detecting moderate to severe fracture. This model could serve as a component of an automated opportunistic screening tool that processes radiographs for a large healthcare system with few false positives. As a result, spinal OCFs could be detected earlier, facilitating further diagnostic workup and earlier treatment to improve quality of life and to decrease mortality and morbidity.

Appendix

Recall that when training the GoogLeNet model, we used random search to tune four hyper-parameters. In Table A.1, we present the definition, the optimal value, and the search range of each of these four hyper-parameters. The initial learning rate was searched on the logarithmic scale, while the other three hyper-parameters were searched on a linear scale.

References:

[1]. Looker AC, Borrud LG, Dawson-Hughes B, Shepherd JA, Wright NC. Osteoporosis or low bone mass at the femur neck or lumbar spine in older adults, United States, 2005–2008. NCHS Data Brief 2012;93:1–8.

[2]. Kanis JA, on behalf of the World Health Organization Scientific Group (2007). Assessment of osteoporosis at the primary health-care level. Technical Report WHO Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK.

[3]. Hodsman AB, Leslie WD, Tsang JF, Gamble GD. 10-year probability of recurrent fractures following wrist and other osteoporotic fractures in a large clinical cohort: an analysis from the Manitoba Bone Density Program. JAMA Internal Medicine 2008;168(20):2261–7.

[4]. Roux S, Cabana F, Carrier N, Beaulieu M, April PM, Beaulieu MC, Boire G. The World Health Organization Fracture Risk Assessment Tool (FRAX) underestimates incident and recurrent fractures in consecutive patients with fragility fractures. J Clin Endocrinol Metab 2014;99(7):2400–8. [PubMed: 24780062]

[5]. Robinson CM, Royds M, Abraham A, McQueen MM, Court-Brown CM, Christie J. Refractures in patients at least forty-five years old: a prospective analysis of twenty-two thousand and sixty patients. J Bone Joint Surg Am 2002;84(9):1528–33. [PubMed: 12208908]

[6]. Center JR, Nguyen TV, Schneider D, Sambrook PN, Eisman JA. Mortality after all major types of osteoporotic fracture in men and women: an observational study. The Lancet 1999;353(9156):878–82.

[7]. Meadows ES, Whangbo A, McQuarrie N, Gilra N, Mitchell BD, Mershon JL. Compliance with mammography and bone mineral density screening in women at least 50 years old. Menopause 2011;18(7):794–801. [PubMed: 21505373]

[8]. Jain S, Bilori B, Gupta A, Spanos P, Singh M. Are men at high risk for osteoporosis underscreened? A quality improvement project. Perm J 2016;20(1):60–4. [PubMed: 26824964]

[9]. King AB, Fiorentino DM. Medicare payment cuts for osteoporosis testing reduced use despite tests' benefit in reducing fractures. Health Aff 2011;30(12):2362–70.

[10]. Pickhardt PJ, Pooler BD, Lauder T, del Rio AM, Bruce RJ, Binkley N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. Ann Intern Med 2013;158(8):588–95. [PubMed: 23588747]

[11]. Anderson PA, Polly DW, Binkley NC, Pickhardt PJ. Clinical use of opportunistic computed tomography screening for osteoporosis. JBJS 2018;100(23):2073–81.

[12]. Alacreu E, Moratal D, Arana E. Opportunistic screening for osteoporosis by routine CT in Southern Europe. Osteoporosis International 2017;28(3):983–90. [PubMed: 28108802]

[13]. Li YL, Wong KH, Law MW, Fang BX, Lau VW, Vardhanabuti VV, Lee VK, Cheng AK, Ho WY, Lam WW. Opportunistic screening for osteoporosis in abdominal computed tomography for Chinese population. Archives of Osteoporosis 2018;13(1):1–7.

[14]. Cheng X, Zhao K, Zha X, Du X, Li Y, Chen S, Wu Y, Li S, Lu Y, Zhang Y, Xiao X. Opportunistic screening using low-dose CT and the prevalence of osteoporosis in China: a nationwide, multicenter study. Journal of Bone and Mineral Research 2021:36(3):427–35. [PubMed: 33145809]

[15]. Fang Y, Li W, Chen X, Chen K, Kang H, Yu P, Zhang R, Liao J, Hong G, Li S. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. European Radiology 2021;31(4):1831–42. [PubMed: 33001308]

[16]. Nam KH, Seo I, Kim DH, Lee JI, Choi BK, Han IH. Machine learning model to predict osteoporotic spine with hounsfield units on lumbar computed tomography. Journal of Korean Neurosurgical Society 2019;62(4):442–9. [PubMed: 31290297]

[17]. Löffler MT, Jacob A, Scharr A, Sollmann N, Burian E, El Husseini M, Sekuboyina A, Tetteh G, Zimmer C, Gempt J, Baum T. Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA. European Radiology 2021;31:6069–77. [PubMed: 33507353]

[18]. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Prediction of bone mineral density from computed tomography: application of deep learning with a convolutional neural network. European radiology 2020;30:3549–57. [PubMed: 32060712]

[19]. Bar A, Wolf L, Amitai OB, Toledano E, Elnekave E. Compression fractures detection on CT. In: Proceedings of SPIE Medical Imaging: Computer-Aided Diagnosis, Orlando, FL. International Society for Optics and Photonics, 2017; 1013440.

[20]. Yilmaz EB, Buerger C, Fricke T, Sagar MM, Peña J, Lorenz C, Glüer CC, Meyer C. Automated Deep Learning-Based Detection of Osteoporotic Fractures in CT Images. In: Proceedings of Machine Learning in Medical Imaging, Strasbourg, France Cham, Switzerland: Springer, 2021; 376–85.

[21]. Husseini M, Sekuboyina A, Bayat A, Menze BH, Loeffler M, Kirschke JS. Conditioned variational auto-encoder for detecting osteoporotic vertebral fractures. In: Proceedings of the International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, Granada, Spain Cham, Switzerland: Springer, 2019; 29–38.

[22]. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Computers in Biology and Medicine 2018;98:8–15. [PubMed: 29758455]

[23]. Lee S, Choe EK, Kang HY, Yoon JW, Kim HS. The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population. Skeletal Radiology 2020;49(4):613–8. [PubMed: 31760458]

[24]. Zhang B, Yu K, Ning Z, Wang K, Dong Y, Liu X, Liu S, Wang J, Zhu C, Yu Q, Duan Y. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. Bone 2020;140:115561. [PubMed: 32730939]

[25]. Murata K, Endo K, Aihara T, Suzuki H, Sawaji Y, Matsuoka Y, Nishimura H, Takamatsu T, Konishi T, Maekawa A, Yamauchi H. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. Scientific Reports 2020;10(1):1–8. [PubMed: 31913322]

[26]. Chou PH, Jou TH, Wu HT, Yao YC, Lin HH, Chang MC, Wang ST, Lu HH, Chen HH. Ground truth generalizability affects performance of the artificial intelligence model in automated vertebral fracture detection on plain lateral radiographs of the spine. The Spine Journal 2021.

[27]. IMV reports general X-ray procedures growing at 5.5% per year, as number of installed X-ray units declines CISION PRWeb. https://www.prweb.com/releases/2011/2/prweb8127064.htm. Accessed October 22, 2020.

[28]. Bolotin HH. DXA in vivo BMD methodology: an erroneous and misleading research and clinical gauge of bone mineral status, bone fragility, and bone remodelling. Bone 2007;41(1):138–54. [PubMed: 17481978]

[29]. Kim TY, Schafer AL. Variability in DXA reporting and other challenges in osteoporosis evaluation. JAMA Internal Medicine 2016;176(3):393–5. [PubMed: 26746871]

[30]. Carberry GA, Pooler BD, Binkley N, Lauder TB, Bruce RJ, Pickhardt PJ. Unreported vertebral body compression fractures at abdominal multidetector CT. Radiology 2013;268(1):120–6. [PubMed: 23449956]

[31]. Khan S, Rahmani H, Shah SA, Bennamoun M. A Guide to Convolutional Neural Networks for Computer Vision. Morgan & Claypool, 2018.

[32]. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA Washington, D.C.: IEEE Computer Society, 2015; 1–9.

[33]. Bianco S, Cadene R, Celona L, Napoletano P. Benchmark analysis of representative deep neural network architectures. IEEE Access 2018;6:64270–7.

[34]. Orwoll E, Blank JB, Barrett-Connor E, Cauley J, Cummings S, Ensrud K, Lewis C, Cawthon PM, Marcus R, Marshall LM, McGowan J, Phipps K, Sherman S, Stefanick ML, Stone K. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study-a large observational study of the determinants of fracture in older men. Contemp Clin Trials 2005;26(5):569–85. [PubMed: 16084776]

[35]. Cawthon PM, Haslam J, Fullman R, Peters KW, Black D, Ensrud KE, Cummings SR, Orwoll ES, Barrett-Connor E, Marshall L, Steiger P, Schousboe JT, Osteoporotic Fractures in Men (MrOS) Research Group. Methods and reliability of radiographic vertebral fracture detection in older men: the osteoporotic fractures in men study. Bone 2014;67:152–5. [PubMed: 25003811]

[36]. Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res 1993;8(9):1137–48. [PubMed: 8237484]

[37]. Gonzalez RC, Woods RE. Digital Image Processing 4th ed. New York, NY: Pearson, 2018.

[38]. imgaug: Read the Docs https://imgaug.readthedocs.io/en/latest/source/api_imgaug.html. Updated 2020. Accessed September 7, 2020.

[39]. TensorFlow https://www.tensorflow.org. Accessed July 25, 2020.

[40]. Silberman N, Guadarrama S. TF-Slim: a high level library to define complex models in TensorFlow Google AI Blog. https://ai.googleblog.com/2016/08/tf-slim-high-level-library-to-define.html. Published August 30, 2016. Accessed July 25, 2020.

[41]. GoogLeNet-Inception. GitHub https://github.com/conan7882/GoogLeNet-Inception. Accessed February 5, 2022.

[42]. TensorFlow Model Garden. GitHub https://github.com/tensorflow/models. Updated July 24, 2020. Accessed July 25, 2020.

[43]. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF. ImageNet large scale visual recognition challenge. IJCV 2015;115:211–52.

[44]. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE international conference on computer vision, Las Condes, Chile Washington, D.C.: IEEE Computer Society, 2015; 1026–34.

[45]. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA New York, NY: Association for Computing Machinery, 2015.

[46]. tf.nn.weighted_cross_entropy_with_logits. TensorFlow https://www.tensorflow.org/api_docs/python/tf/nn/weighted_cross_entropy_with_logits. Accessed September 7, 2020.

[47]. Goodfellow I, Bengio Y, Courville A. Deep Learning Cambridge, MA: MIT press, 2016.

[48]. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine learning, Pittsburgh, PA New York, NY: Association for Computing Machinery, 2006; 233–40.

[49]. Lentle BC, Berger C, Probyn L, Brown JP, Langsetmo L, Fine B, Lian K, Shergill AK, Trollip J, Jackson S, Leslie WD, Prior JC, Kaiser SM, Hanley DA, Adachi JD, Towheed T, Davison KS, Cheung AM, Goltzman D, CaMos Research Group. Comparative analysis of the radiology of osteoporotic vertebral fractures in women and men: cross-sectional and longitudinal observations from the Canadian Multicentre Osteoporosis study (CaMos). J Bone Miner Res 2018;33(4):569–79. [PubMed: 28722766]

[50]. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV Washington, D.C.: IEEE Computer Society, 2016; 770–8.

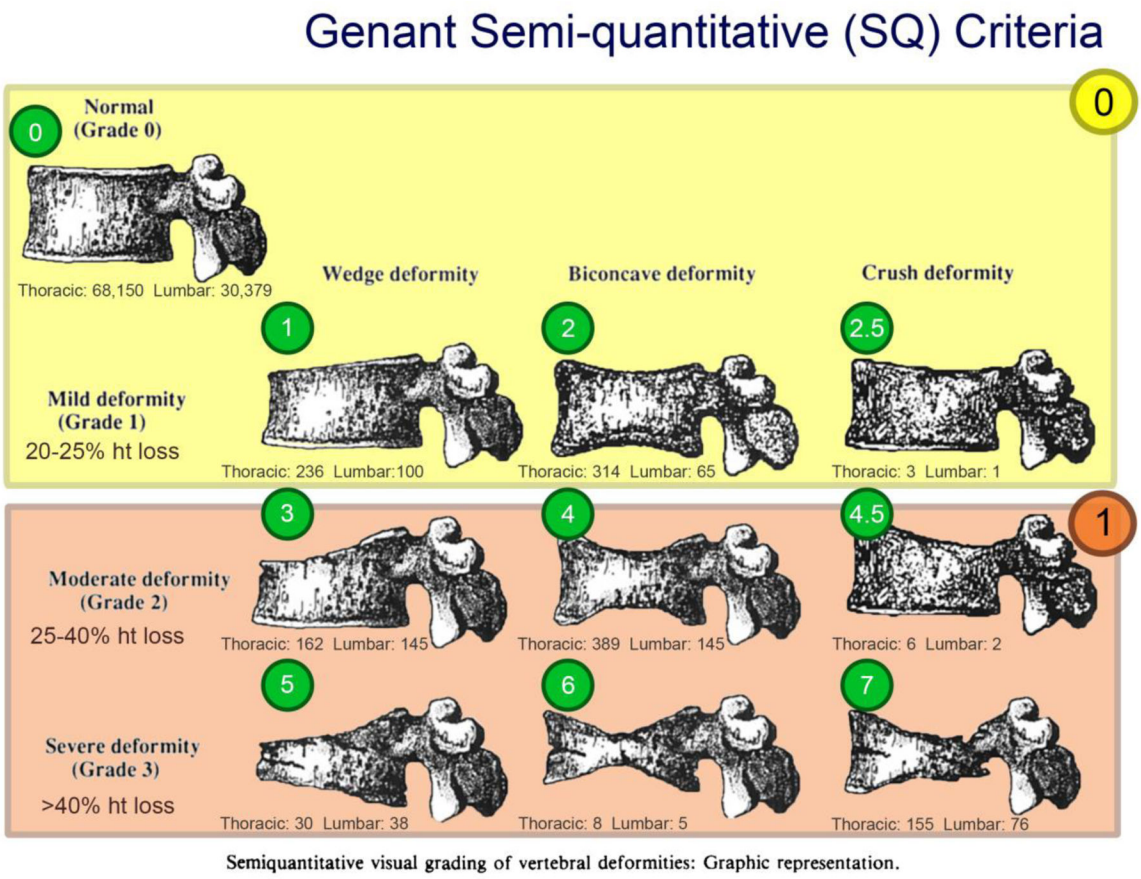


Figure 1:

A graphic representation of the Genant semiquantitative (SQ) osteoporotic fracture classification criteria. This approach to facture classification uses nine fracture classes and one normal class. Fractures are graded by the degree of height loss (mild, moderate, or severe) and whether the vertebral body height loss is predominantly anterior, posterior, or central. The MrOS dataset assigns the 10 classes the numerical labels: 0, 1, 2, 2.5, 3, 4, 4.5, 5, 6, and 7 (green bubbles). The original Genant criteria were modified slightly by the MrOS study to include the requirement for depression of the endplate to be present for the “Mild deformity” row [35]. This system was simplified into two classes: label 0 (yellow) representing a normal or possible, mild deformity, and label 1 (orange) representing a moderate to severe deformity. Adapted from: Genant HK, Wu CY, van Kujik C, Nevitt MC: Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. Sep; 1148, 1993. Fig. 1: Semiquantitative visual grading of vertebral deformities: Graphic representation. © 1993 American Society for Bone and Mineral Research.

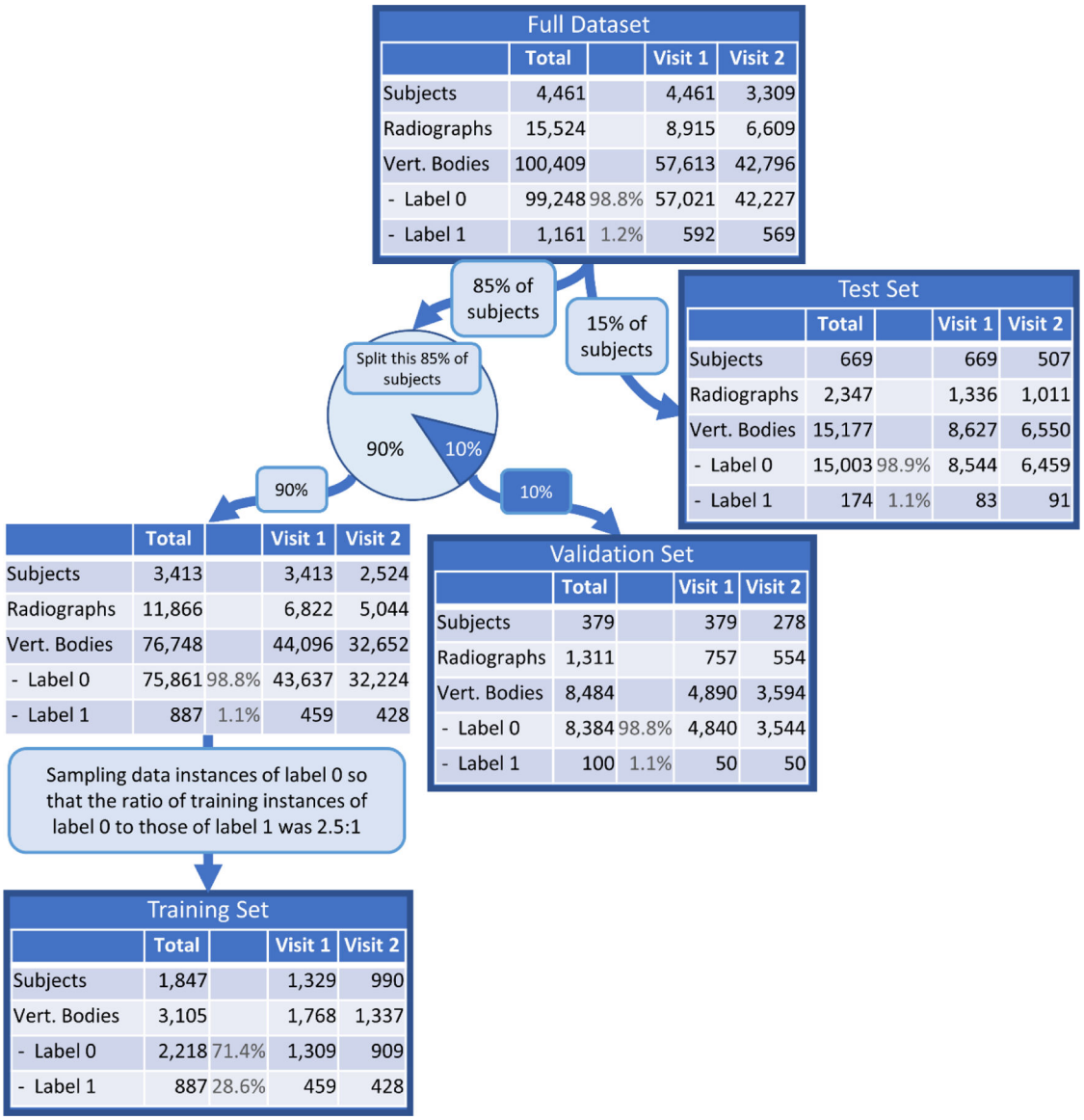


Figure 2:

The MrOS dataset was divided into the test, validation, and training sets by subject. A thoracic and lumbar radiograph was obtained at both the first clinical visit (Visit 1) and the follow-up clinical visit (Visit 2). Since the radiographs of the same subject have some commonality, datasets were divided on a subject basis. To reduce the data imbalance degree in the training set, instances of label 0 (normal/possible/mild deformity) were subsampled to the ratio of 2.5:1 (label 0 to label 1) in order to better balance the cases in the two classes.

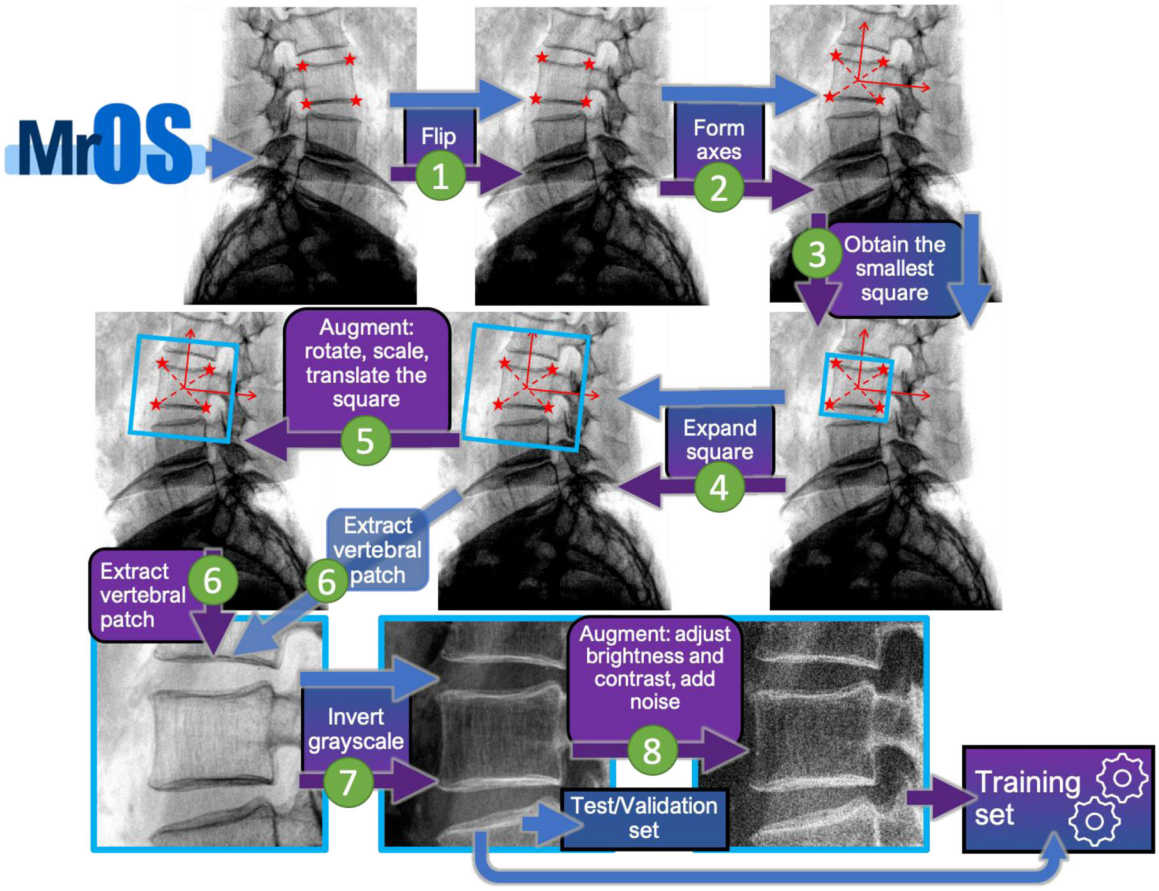


Figure 3:

This process of generating a vertebral patch was performed for each vertebral body labeled in a radiograph using the four corner points indicated by red stars. The blue and purple arrows demonstrate the creation of the vertebral patches without and with the augmentation steps, respectively. The vertebral patches in the validation and test sets should not be augmented. Both the raw and augmented vertebral patches were included in the training set. The steps are: 1) flip the radiograph horizontally to conform to the convention that the subject faces left; 2) form two coordinate axes with the x-axis bisecting the angle between the two diagonals connecting the four corner points; 3) obtain the smallest square that bounds the four corner points with edges parallel to the corresponding coordinate axes; 4) expand the square from its center to increase the area by four times, preventing cutoff of part of the vertebral body and providing surrounding image context; 5) augment the vertebral patch by scaling, rotating, and translating the square; 6) extract the square as a vertebral patch; 7) invert the grayscale if the bones are darker than the background; 8) augment the vertebral patch by changing the contrast and brightness and adding Gaussian noise to the vertebral patch; 9) resize the vertebral patch to 224×224 pixels and normalize each pixel value by subtracting the mean and then dividing by the standard deviation [31].

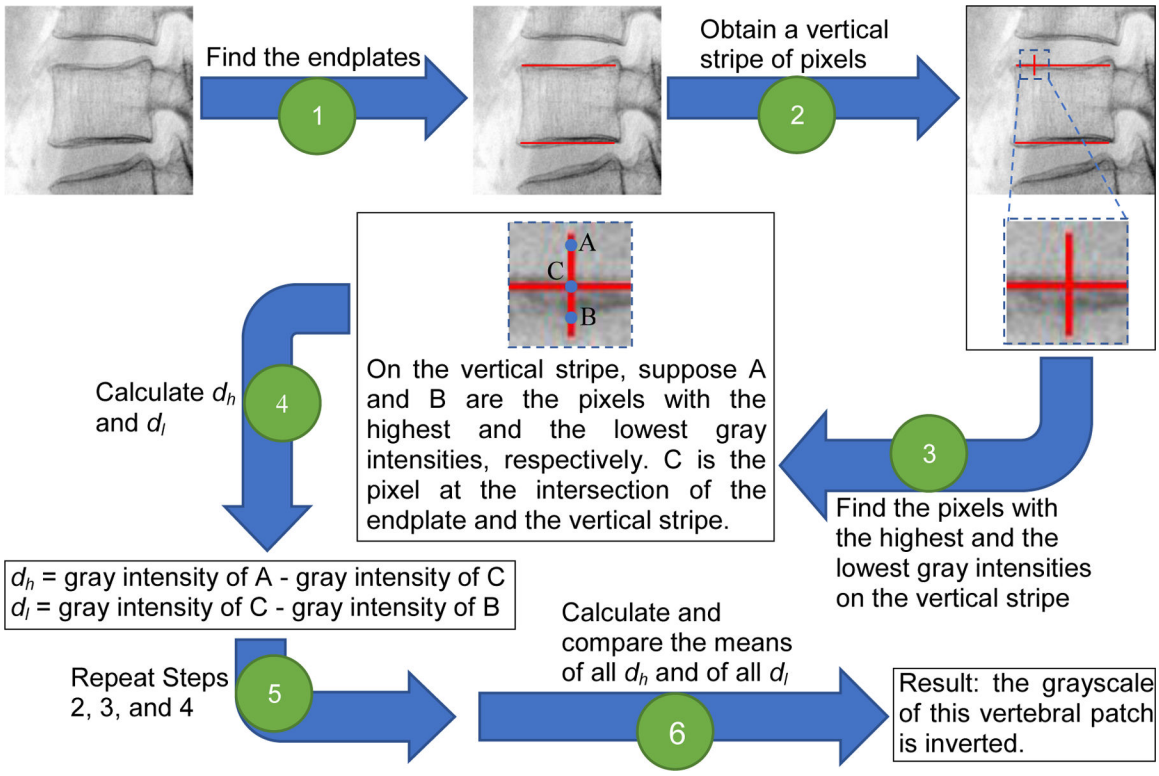


Figure 4:

The process to determine whether the grayscale of a vertebral patch is inverted. The steps are: 1) find the endplates using Sobel operator and hysteresis thresholding; 2) on the endplate, obtain a vertical strip of pixels whose midpoint is the pixel on the endplate; 3) on the vertical stripe, find the pixels with the highest and the lowest gray intensities; 4) calculate d_h and d_l; 5) traverse the pixels on the endplates and repeat Steps 2, 3, and 4 to get all d_h and all d_l; 6) calculate the means of all d_h and of all d_l, respectively, and compare them to determine whether the grayscale of the vertebral patch is inverted. If the mean of all d_l is < the mean of all d_h, the grayscale of the vertebral patch is inverted; otherwise, the vertebral patch is standard.

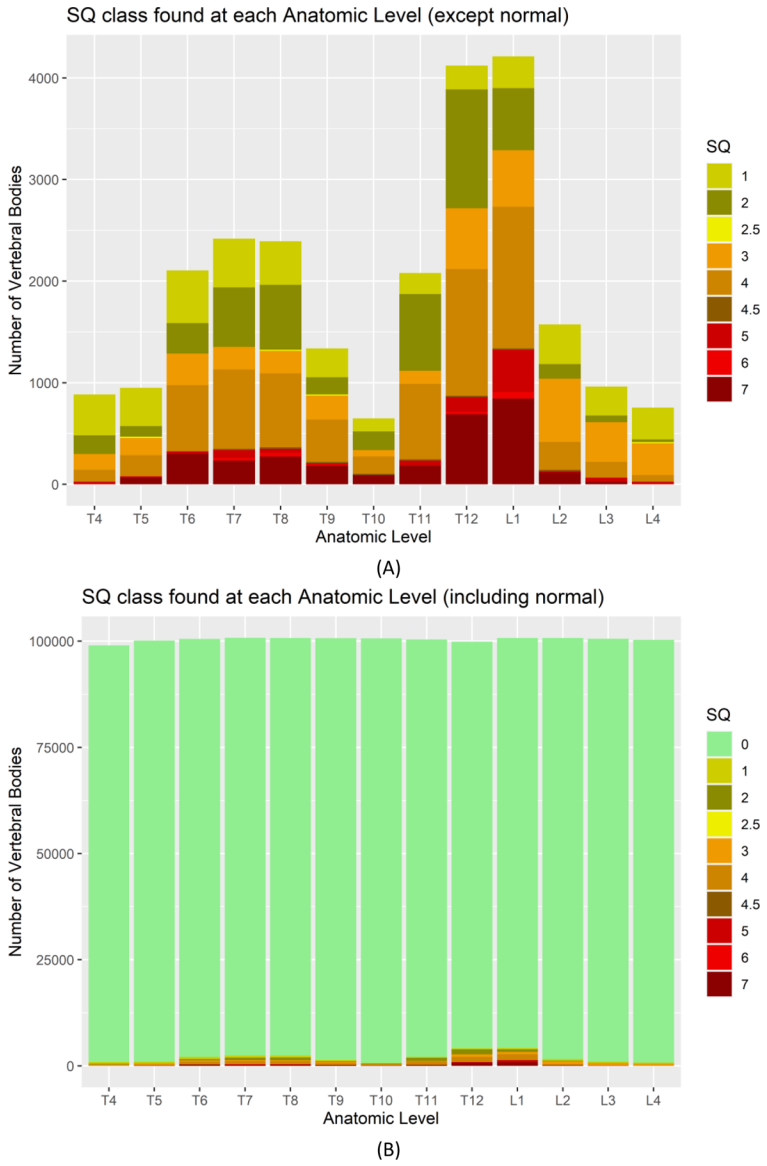
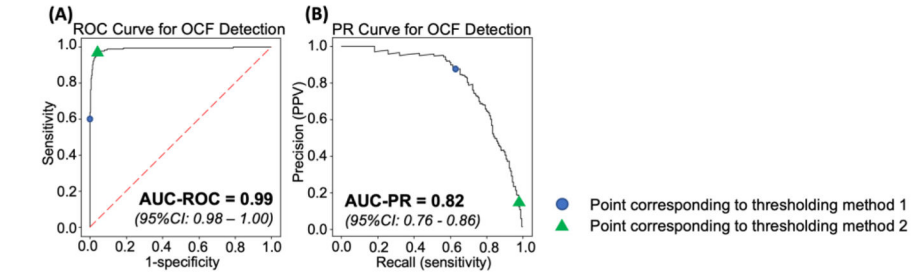


Figure 5:

In the entire MrOS dataset, the number of vertebral bodies of each SQ class at each anatomic level of the spine (A) excluding and (B) including the normal class. Each digit in the figure's legend represents an SQ class shown in Figure 1 (green bubbles).



(C)

Thresholding method	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	FDR (%)	F_1	Accuracy (%)
Method 1: Prioritize PPV and balance it with sensitivity	59.8% [51.9%-67.1%]	99.9% [99.9%-100.0%]	91.2% [86.1%-96.1%]	99.5% [99.4%-99.6%]	8.8% [3.9%-13.9%]	0.72 [0.66-0.78]	99.5% [99.4%-99.6%]
Method 2: Maximize Youden's J statistic	97.7% [95.3%-99.5%]	95.1% [94.7%-95.4%]	18.8% [16.3%-21.3%]	100.0% [99.9%-100.0%]	81.2% [78.7%-83.7%]	0.31 [0.28-0.35]	95.1% [94.8%-95.5%]

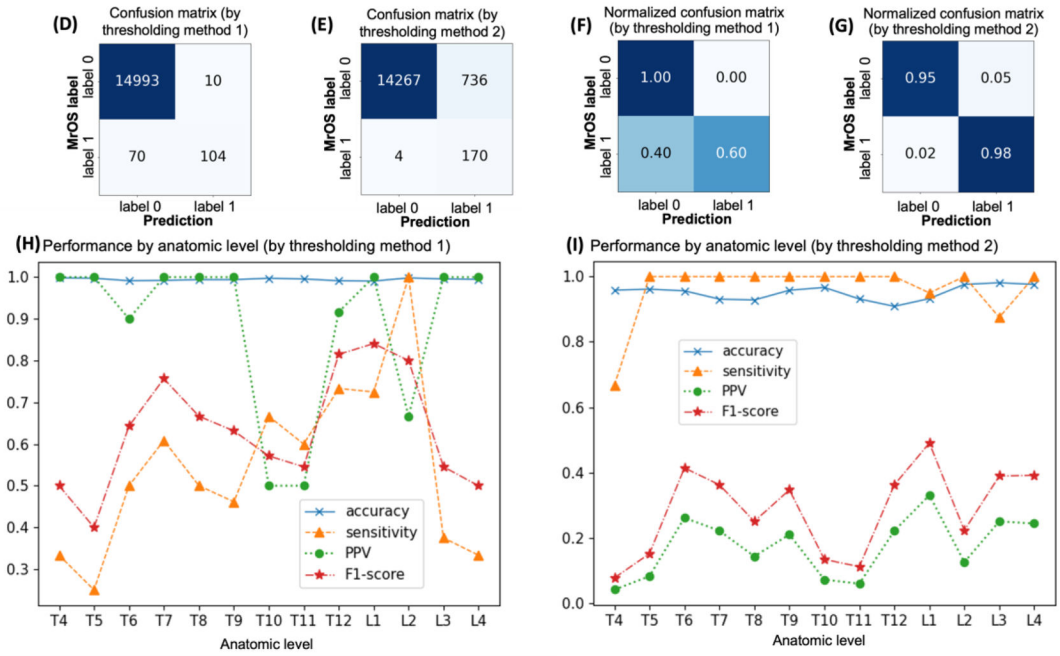


Figure 6:

On the test set, the final deep learning model achieved an AUC-ROC of 0.99 (A) and an AUC-PR of 0.82 (B) with the associated 95% confidence intervals (CIs). Two thresholding methods are used. Their corresponding sensitivities, specificities, PPVs, NPVs, FDRs, F_1 scores, and accuracies are shown in (C). The values in each pair of brackets in (C) represent the 95% CI. Thresholding method 1 provides a favorable PPV and a favorable FDR for large volume screening. Thresholding method 2 balances sensitivity and specificity by optimizing Youden's J statistics. The confusion matrices generated using thresholding methods 1 and 2 are shown in (D) and (E), respectively. For each confusion matrix, to normalize the values in it, each element in each row of it is divided by the sum of the elements in the row. The normalized confusion matrices are shown in (F) and (G). With thresholding methods 1 and 2 used, (H) and (I) show the accuracy, sensitivity, PPV, and F_1 score by each anatomic level of the spine. Note that the number of fractured vertebral bodies at T10 and T11 are three

and five, respectively. Because these numbers are small, there is limited statistical power to evaluate the model at these two anatomic levels (T10 and T11).

Table 1:

Demographic information of the subjects in each of the entire, training, validation, and test sets from the MrOS dataset. Due to a large imbalance between label 0 and label 1 in the training set, label 0 was downsampled to reduce the imbalance between these two classes. Both the characteristics of the subjects in the training set before downsampling and those in the training set after downsampling are shown. The mean (standard deviation) of the ages and body mass indices were recorded at the baseline (Visit 1) and the follow-up (Visit 2) visits. Race and ethnicity and the total number of subjects are also provided for each set.

	Training set before downsampling the data instances with label 0	Training set after downsampling the data instances with label 0	Validation set	Test set	Entire dataset
		Mean ± standard deviation			
Age at Visit 1	73.7±5.9	73.7±5.8	74.1±6.2	73.5±5.7	73.7±5.9
Body mass index at Visit 1	27.8±3.9	27.9±4.0	27.2±3.5	27.6±3.5	27.4±3.8
Age at Visit 2	77.8±5.6	77.9±5.6	77.9±5.6	77.5±5.4	77.7±5.6
Body mass index at Visit 2	27.3±3.9	27.3±4.0	27.4±4.0	27.3±3.9	27.3±3.9
		Percentage			
Race/ethnicity					
American Indian or Alaska Native	0.8%	0.7%	1.8%	1.2%	0.9%
Asian	3.2%	2.5%	3.1%	3.7%	3.2%
Black or African American	4.2%	3.0%	5.4%	3.1%	4.2%
Hispanic or Latino	2.0%	2.0%	2.8%	2.2%	2.1%
Native Hawaiian or Other Pacific Islander	0.2%	0.3%	0.8%	0.1%	0.2%
White	89.6%	91.5%	86.1%	89.7%	89.4%
		Number			
Total subjects	5,016	1,874	392	681	6,089

Table 2:

Error analysis of potential confounders that led to incorrect classification of vertebral bodies by the deep learning model. Two neuroradiologists first reviewed a set of 50 random cases to identify the confounders, and then reviewed all 270 random cases to record the presence of each confounder for each case.

	Number of vertebral patches			
Confounder		**Prediction**		***Odds ratio***
		Incorrect	**Correct**	
Substantial noise	present	13	4	*4.43*
	not	107	146	
Part of vertebral patch was cut off and zero padded during extraction because the vertebral body was at the edge of the radiograph	present	6	2	*3.89*
	not	114	148	
Overlying metal object (surgical clip, catheter marker, staples, etc.)	present	1	0	*3.78*
	not	119	150	
Inverted grayscale of the vertebral patch	present	7	5	*1.80*
	not	113	145	
Tape/sticker on vertebral patch	present	3	3	*1.26*
	not	117	147	
Central disk depression (large Schmorl's Node)	present	28	30	*1.22*
	not	92	120	
Heavy parallax artifact (scoliosis or patient positioning)	present	17	21	*1.01*
	not	103	129	
Fractured vertebral body fused to the adjacent vertebral body	present	15	20	*0.93*
	not	105	130	
Significant disk space calcification	present	14	21	*0.81*
	not	106	129	
Strong contrast gradient superimposed over the vertebral body (usually from the diaphragm or iliac crest)	present	25	37	*0.80*
	not	95	113	
Writing on the vertebral patch	present	11	18	*0.74*
	not	109	132	
Prominent overlying lines from ribs or linear structures projecting over the vertebral body	present	45	86	*0.45*
	not	75	64	
Poor definition of the margins of the vertebral body	present	20	48	*0.43*
	not	100	102	
Point placement wrong, resulting in a distorted vertebral patch	present	0	1	*0.41*
	not	120	149	
Very bulky osteophytes	present	20	50	*0.40*
	not	100	100	

Table 3:

The GoogLeNet model outputs a predicted confidence for OCF for each vertebral patch, to which a variable threshold can be applied to change the percentage of data instances that are classified as label 1, resulting in a spectrum of sensitivities, specificities, PPVs, NPVs, FDRs, F_1 scores, and accuracies.

Cutoff percentage	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	FDR (%)	F_1 score	Accuracy (%)
0.25	21.3	100.0	97.4	99.1	2.6	0.35	99.1
0.50	42.0	100.0	96.1	99.3	3.9	0.58	99.3
0.75	59.8	99.9	91.2	99.5	8.8	0.72	99.5
1.00	70.1	99.8	80.3	99.7	19.7	0.75	99.5
1.25	76.4	99.6	70.0	99.7	30.0	0.73	99.4
1.50	81.6	99.4	62.3	99.8	37.7	0.71	99.2
2.00	85.6	99.0	49.0	99.8	51.0	0.62	98.8
3.00	92.5	98.0	35.3	99.9	64.7	0.51	98.0
4.00	94.3	97.0	27.0	99.9	73.0	0.42	97.0
6.00	97.7	95.1	18.7	100.0	81.3	0.31	95.1
10.00	98.3	91.0	11.3	100.0	88.7	0.20	91.1

Table 4:

The GoogLeNet model's performance measures on the validation set, the training set without the augmented vertebral patches, and the training set with the augmented vertebral patches. For each set, the classification threshold on the predicted probability of having label 1 was set to maximize the Youden's J statistics.

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	FDR (%)	F_1 score	Accuracy (%)
Validation set	100.0	95.8	22.2	100.0	77.8	0.36	95.9
Training set without augmented vertebral patches	100.0	99.3	98.2	100.0	1.8	0.99	99.5
Training set with augmented vertebral patches	99.9	98.6	96.6	100.0	3.4	0.98	99.0

Table A.1:

The four hyper-parameters that were tuned by random search over 1,500 rounds.

Hyper-parameter	Description	Optimal value	Search range
Initial learning rate	The learning rate when model training starts.	6.95×10^{-4}	$[10^{-6}, 10^{-2}]$
Learning rate decay	The value by which the learning rate was divided at the end of each epoch.	8.53	[1.0, 10.0]
Dropout rate	For each unit in the fully connected layer before the output layer, the probability of dropping it.	0.25	[0.0, 1.0]
pos_weight	The factor that controls the false positives' weight in the weighted cross entropy loss function [25].	0.14	[0.0, 1.0]