Computers in Biology and Medicine 98 (2018) 8–15

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans

Naofumi Tomita [a], Yvonne Y. Cheung [b], Saeed Hassanpour [a,c,d,*]

[a] Biomedical Data Science Department, Dartmouth College, Hanover, NH 03755, USA
[b] Radiology Department, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, USA
[c] Epidemiology Department, Dartmouth College, Hanover, NH 03755, USA
[d] Computer Science Department, Dartmouth College, Hanover, NH 03755, USA

ARTICLE INFO

Keywords:
Deep convolutional neural network
Osteoporotic vertebral fracture
Imaging informatics
Volumetric CT classification
Recurrent neural network

ABSTRACT

Osteoporotic vertebral fractures (OVFs) are prevalent in older adults and are associated with substantial personal suffering and socio-economic burden. Early diagnosis and treatment of OVFs are critical to prevent further fractures and morbidity. However, OVFs are often under-diagnosed and under-reported in computed tomography (CT) exams as they can be asymptomatic at an early stage. In this paper, we present and evaluate an automatic system that can detect incidental OVFs in chest, abdomen, and pelvis CT examinations at the level of practicing radiologists. Our OVF detection system leverages a deep convolutional neural network (CNN) to extract radiological features from each slice in a CT scan. These extracted features are processed through a feature aggregation module to make the final diagnosis for the full CT scan. In this work, we explored different methods for this feature aggregation, including the use of a long short-term memory (LSTM) network. We trained and evaluated our system on 1432 CT scans, comprised of 10,546 two-dimensional (2D) images in sagittal view. Our system achieved an accuracy of 89.2% and an F1 score of 90.8% based on our evaluation on a held-out test set of 129 CT scans, which were established as reference standards through standard semiquantitative and quantitative methods. The results of our system matched the performance of practicing radiologists on this test set in real-world clinical circumstances. We expect the proposed system will assist and improve OVF diagnosis in clinical settings by pre-screening routine CT examinations and flagging suspicious cases prior to review by radiologists.

1. Introduction

Osteoporosis, a chronic progressive bone disorder related to loss of bone density and quality, affects 10.2 million Americans [1] and 56.2 million people worldwide [2]. Weakened bone leads to fragility fractures that are associated with loss of independence and a decrease in the quality of life. If osteoporosis is detected early, it can be effectively treated to reduce future fractures and morbidity. However, before the onset of symptomatic fractures, osteoporosis is frequently silent, resulting in under-diagnosis and under-treatment.

Osteoporotic vertebral fracture or OVF, a marker of osteoporosis, is the most common type of osteoporotic fracture. The prevalence of OVF is high in older adults, reaching 40% by the age of 80 [3]. Nevertheless, under-reporting of incidental OVFs remains common with 84% of OVFs not reported in computed tomography (CT) exams in one study [4]. The under-reporting of incidental OVFs in routine CT exams is attributed to radiologists' inattention to the sagittal views, the absence of clinical symptoms [5–8], and a lack of awareness regarding the clinical importance of asymptomatic OVFs [9].

There are new opportunistic approaches to screen for osteoporosis [9–15]. These options are opportunistic because they rely on CT examinations performed for indications not related to the spine. As a result, the radiologist could be the first to suspect osteoporosis based on the imaging findings. Furthermore, these new screening approaches are efficient, because they do not require extra imaging time or radiation dose.

Recent advancement of machine learning allows automatic diagnosis of various conditions on radiology exams [12,16,17]. Such automatic diagnosis has many benefits. For instance, radiologists no longer need to perform the tedious task of screening for incidental findings, and the saved time allows them to interact more with patients and health providers. Furthermore, these automatic diagnostic tools can address the lack of access to expert radiologists in rural, small, or poor communities.

* Corresponding author. Dartmouth College, One Medical Center Drive, HB 7261, Lebanon, NH 03756, USA.
E-mail address: Saeed.Hassanpour@dartmouth.edu (S. Hassanpour).

https://doi.org/10.1016/j.compbiomed.2018.05.011
Received 20 February 2018; Received in revised form 4 May 2018; Accepted 5 May 2018
0010-4825/© 2018 Elsevier Ltd. All rights reserved.

Particularly, embracing machine learning technology for detecting OVFs can improve early diagnosis of osteoporosis, initiate treatment, and predict future fragility fractures. As a result, a successful OVF detection system could potentially decrease the socio-economic burden of osteoporosis [18].

Previous work on automatic OVF detection relied on multiple and fragmented steps on each vertebra [12,19–22]. These approaches were inefficient for this detection task because they required vertebra segmentation and calculations of height loss ratio on individual vertebral bodies. In this study, we developed and evaluated an automatic detection system for OVFs, based on an end-to-end deep learning model, which does not require multiple segmentation and analysis steps for each vertebra. In our proposed system, we leverage a convolutional neural network (CNN) to extract radiological features from chest, abdomen, and pelvis CT exams. The resulting sequences of features were then aggregated through a sequence classifier to predict the presence of a vertebral fracture on a CT scan. In this study, we aimed to identify an effective feature aggregation methodology to detect OVFs and to compare the performance of this model to the radiologists' performance in our institution. The details of the proposed system and its evaluation are described in the rest of the paper.

2. Methods

2.1. System overview

For our OVF detection system, we explored different feature aggregation methods using a recurrent neural network (RNN)-based model and three rule-based approaches in combination with a CNN feature extractor. Fig. 1 shows the overview of our RNN-based OVF detection system. This system has two major components: (1) a CNN-based feature extraction module; and (2) an RNN module to aggregate the extracted features and make the final diagnosis. For processing and extracting features from two-dimensional (2D) CT slices, we used a deep residual network (ResNet), which is one of the state-of-the-art architectures and achieves the most compelling accuracy for image detection [23]. In our RNN-based feature aggregation, the extracted features from each slice were fed to a long short-term memory (LSTM) network [24]. Among different RNN architectures, LSTMs are state-of-the-art and are often utilized to aggregate and analyze temporal or sequential data such as videos or text. For our task, LSTM was used to aggregate a flexible number of slices from a CT scan to make the final diagnosis.

In addition to the LSTM-based aggregation method, we explored three rule-based alternatives for feature aggregation. These rule-based approaches were directly built into the CNN component and aggregate the confidence scores from each CT slice through (1) taking the maximum, (2) averaging, and (3) voting operations to make the final diagnosis. We compared the performance of our LSTM-based and rule-based approaches through our evaluation on an independent test set.

2.2. Feature extraction network

An accurate clinical diagnosis requires a set of well-established features that represents a radiology examination. In this work, we replace the traditional feature extraction process by radiologists with a CNN to extract high-level features in 2D CT slices. A CNN with an appropriate capacity of parameters and number of layers can represent characteristic features in complex datasets. In our system, we use a ResNet architecture with 34 layers (i.e., ResNet34 architecture) to extract features from each slice in a CT scan.

The ResNet34 architecture consists of 16 two-layer residual blocks, and a total of 33 convolutional layers and one fully connected (FC) layer. In our preliminary experiments, ResNet34 achieved similar results compared to deeper ResNet architectures with shorter training time. The configuration of this feature extraction network was consistent in our all experiments with feature aggregation methods. To adapt this architecture for a binary classification task and to reduce the dimensionality of internal feature representations, we replaced the FC layer of the original ResNet34 architecture with two FC layers, each followed by a batch normalization layer and a rectified linear unit (ReLU) function [25]. The first FC layer reduces the dimensionality of features from 512 to 32, and the second FC layer projects the 32-dimensional vector to a scalar value. This scalar value is normalized by a sigmoid function, $f(x) = \frac{1}{1+e^{-x}}$, to produce a probability value $\hat{y}\varepsilon[0,1]$ as the CNN output. The CNN is trained to make a slice-level classification by minimizing the binary cross-entropy loss between the predicted probability and the reference

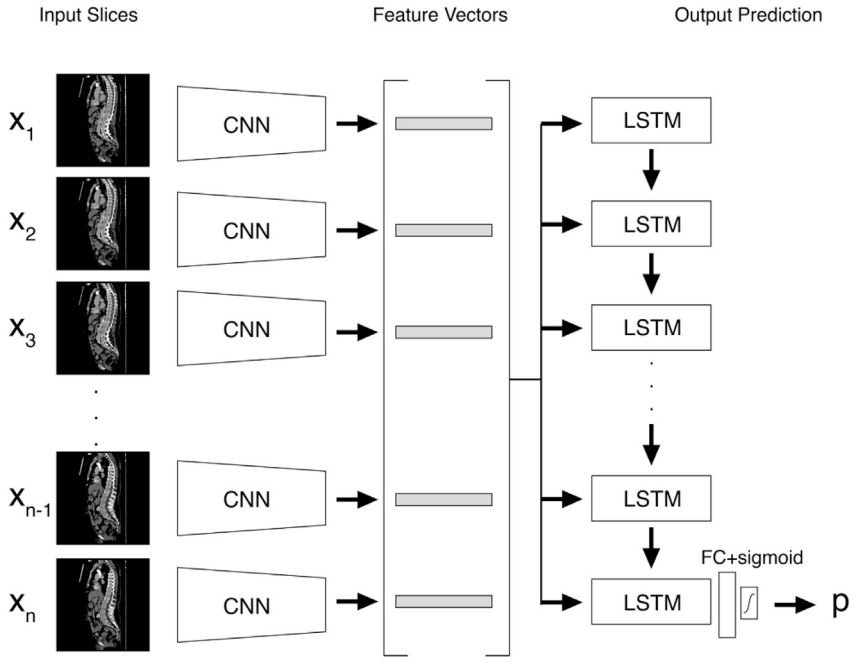

Fig. 1. An Overview of our CNN/LSTM OVF detection system using CNN for feature extraction and LSTM for feature aggregation to make the final diagnosis.

probability distribution as described below. Here, y_i represents the target label for the i-th slice in a CT scan and the label of a CT volume is transferred to all its slices.

$$loss(\hat{y}, y) = -\frac{1}{n}\sum_i \left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$$

2.3. Feature aggregation

The sequence of features from each CT slice is utilized by our feature aggregation module to make a patient-level diagnosis. Since each CT examination has a different number of slices, our feature aggregator must be able to process an arbitrary number of CT slices. In this work, we developed and evaluated four different approaches to aggregate and classify an arbitrary sequence of feature vectors for a CT scan.

Our RNN-based model (CNN/RNN) is a one-layer LSTM network with 256 hidden units followed by an FC layer and a sigmoid layer. The 32-dimensional feature vectors extracted from each CT slice are fed to the network in a sequence, from the first to the last slice. The hidden layer output is a 256-dimensional vector. The output at the last slice is used for the final prediction through FC and sigmoid layers (Fig. 1). The final sigmoid layer produces a single normalized value between 0 and 1 as the probability of OVF on a CT scan (i.e., the confidence score). We set the threshold to 0.5 for these confidence scores for the final binary classification, which is middle point probability threshold to differentiate negative and positive predictions.

Our rule-based methods use slice-level predictions from the last FC layer of the ResNet34 model and aggregate sequential predictions through (1) taking the maximum, (2) averaging, and (3) voting. Given a CT scan with n slices and their corresponding predictions from the last FC layer of ResNet34, $\hat{Y}_i = [\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{in}]$, the description of these rule-based aggregation methods are as follows:

CNN/Max: Taking the Maximum
1 if $sup(\hat{Y}_i) > 0.5$, 0 otherwise
CNN/Avg: Averaging
1 if $E(\hat{Y}_i) > 0.5$, 0 otherwise
CNN/Vote: Voting
1 if $\frac{\sum_{k=1}^{n} 1_{\hat{y}_{ik} > 0.5}}{n} > \theta$, 0 otherwise

Where $sup(x)$ and $E(x)$ are the supremum function and expected value, respectively, θ is a threshold $\in [0, 1]$; and $1_{\hat{y}_{ik} \geq 0.5}$ is an indicator function, which returns a value of "1" if the confidence score for the kth slice, $\hat{y}_{ik}$, is greater than 0.5, and returns a value of "0" otherwise. In this work, we choose the threshold, θ=0.5, which resulted in the highest accuracy on the validation set.

3. Experiments

3.1. Utilized dataset

The cohort in our study consists of all CT exams performed at our tertiary academic care center at Dartmouth-Hitchcock Medical Center (DHMC; Lebanon, NH) from April 16, 2008, to January 9, 2017, and stored in a Montage Search and Analytics™ server (Montage Healthcare Solutions, Philadelphia, PA). The studies performed at DHMC prior to April 16, 2008, were not available through Montage. We used the Montage search functionality for radiology reports to select all CT exams of the chest, abdomen, and pelvis based on exam codes. We then used the advanced search feature in Montage to select positive and negative cases for OVF according to the following criteria:

1) Positive cases

We used the search term "compression deformity OR compression fracture" in a positive context in CT radiology reports to find potential positive cases of OVF.

2) Negative cases

We used the lack of or negated search term "compression deformity OR compression fracture" in CT radiology reports to find negative cases. This search yielded a large number of cases. Among these, we randomly selected the same number of cases as in the positive group and matched for age and sex.

For these positive and negative cases, we transferred the sagittal reformatted images from the Picture Archiving and Communication System (PACS) (Philips iSite PACS v3.6, Philips Healthcare, Best, Netherlands) to a local encrypted server. The following cases were excluded from image transfer:

1) Incomplete Data

Exams that did not include the entire thoracic and lumbar vertebral bodies in the sagittal reformat.

2) Artifacts and noises

Exams that had significant hardware or motion artifacts.

The details of this data extraction process and the number of exams that were included/excluded in each step are shown in Fig. 2. This data extraction process resulted in 1432 CT scans, 713 positive and 719 negative OVF cases. We randomly split the dataset into 80% training, 10% validation, and 10% test sets. This separation is preserved throughout our experiments.

The training set is used to train our models, while the validation set is used to tune the hyper-parameters for these models. The training and validation sets are labeled according to the positive/negative criteria above, based on radiology reports. Therefore, the training/validation set labels are prone to errors, due to potential missed diagnosis by radiologists or non-standard reporting, which reflects real-world radiology practice. Using these original labels maximizes the practicality of our approach, while it may potentially affect the efficacy of the trained model.

Our test set is used to evaluate the generalizability of our approach on unseen CT exams. To evaluate the performance of our method on the test set, we took extra consideration to identify true-label reference standards for the CT exams in the test set. These reference standards were established through a careful domain-expert radiologist semiquantitative and quantitative re-evaluation, which is described in the Test Set Adjudication section below.

3.2. Data preprocessing

As described in the last section, we collected 713 positive and 719 negative CT scans for this study. Due to differences among patients and CT scanners, the number of slices in a CT scan varies in our dataset; on average, 146.4 $\pm$ 2.3 slices are in each CT scan. The size of each CT scan slice is 512 $\times$ 512 pixels in a singlecolor channel. We converted these CT scans from digital imaging and communications in medicine (DICOM) format to JPEG2000 format with the compression ratio of 10:1, which is visually lossless [26]. Also, to normalize the intensity values, a pixel-wise mean subtraction was performed on the CT slices in our dataset. The mean in this normalization was calculated on the training set.

The labels of our training examples are location-agnostic: a positive label only indicates the existence of a vertebral fracture on a CT scan

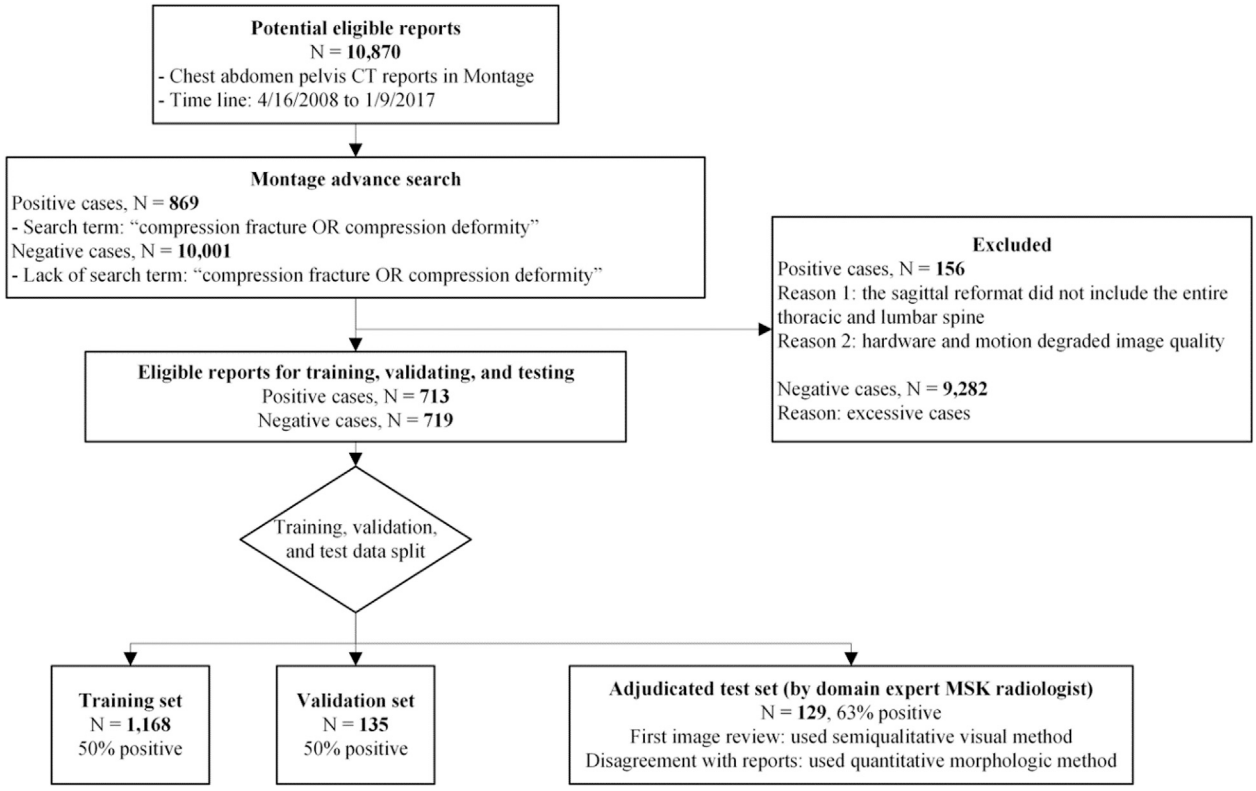

Fig. 2. An overview of our data collection process. In total, 1432 chest, abdomen, and pelvis CT scans were collected for this study. This dataset was partitioned into training, validation, and test sets for the development and evaluation of our system.

without indicating the location of the fracture. Of note, the relevant CT slices to detect vertebral fractures are the ones that contain spinal column anatomy. Through our investigation on the training set, we observed these relevant slices are mostly composed of the 5% middle slices in a typical chest, abdomen, and pelvis CT scan. Therefore, we extracted only the middle 5% of slices of the CT exams in our analysis. Although this 5% criterion is chosen heuristically through visual inspection of our training set, this criterion is flexible for any CT volume in the sagittal view. This extraction narrows down the focus of our method on the slices that are important for OVF detection, and it also reduces the noise introduced by considering slices not showing any parts of the spinal column. In addition, this approach reduces the amount of computation at training and inferencing by only examining the most relevant portion of data. This data preprocessing can reduce the computational cost of our model dramatically (about 20 times) without affecting its accuracy. After this extraction, each sample CT scan contained 6.9 ± 0.1 slices on average.

Data augmentation is an important preprocessing step for deep convolutional neural networks to achieve a generalizable performance based on our relatively small dataset in this study. We used random rotation (rotating an image in the range of $\pm 3°$) and horizontal translation (12 pixels of 0-padding on each side of an image, followed by 512×512 random cropping). Although we tested several other augmentation methods, such as elastic distortions [27], lens distortions [28], and random noise addition, we decided to use only random rotation and horizontal translation in our final configuration, as we did not observe significant improvements by using other data augmentation techniques on our validation set.

To feed an augmented CT slice to a ResNet34 network whose size of input is a $3 \times 224 \times 224$ tensor, we first subsampled a slice of $1 \times 512 \times 512$ into $1 \times 224 \times 224$, and then repeated it in threefold to replicate RGB channels. Our subsampling scheme includes resizing the image to 224×512 and cropping out the central 224×224 patch. We also experimented with two other subsampling schemes (center cropping and resizing). However, our scheme showed a more generalizable performance on our validation set.

3.3. Feature extraction training

A ResNet34 model with two FC layers (see the Feature Extraction Network section) was employed to extract features from slices in CT scans. This model was trained by using the binary cross-entropy loss function and stochastic gradient descent. The second-to-the-last FC layer outputs, 32-dimensional feature vectors, are used as inputs for the subsequent LSTM classifier, while the last FC layer outputs, confidence scores between 0 and 1, are used in our rule-based methods.

We used a ResNet34 model defined as part of torch-vision in the PyTorch framework [29] for training. All parameters throughout this training process are initialized according to He et al. [30] and the training mini-batches are selected randomly. The initial learning rate was 0.01 for the mini-batch size of 192. This learning rate was reduced in half whenever we observed no improvement in the objective loss on the validation set for 30 epochs. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e\text{-}8$ [31]. In addition to data augmentation, weight decay of 1e-4 is utilized to stabilize the training process. We trained our model for 400 epochs. All parameters for the batch normalization layers were frozen after this training.

The training of the ResNet34 feature extraction network took 9 h on a high-performance computer that was equipped with an NVIDIA Titan Xp GPU, an Intel Xeon E5-1650 CPU, and 16 GB of RAM.

3.4. Feature aggregation training

One-layer LSTM followed by a fully connected layer and a sigmoid layer were used as our RNN-based aggregation model. This LSTM network has 256 hidden units, which achieved the best performance among models with different configurations (e.g., 64, 128, 256, 512,

Table 1
The performance of our models for the detection of OVFs on CT scans in comparison to practicing radiologists on our adjudicated test set.

Model	Accuracy (%) (TP + TN)/ (TP + FP + FN + TN)	Precision (%) TP/ (TP + FP)	Sensitivity (%) TP/ (TP + FN)	Specificity (%) TN/ (TN + FP)	F1 score (%) 2TP/ (2 TP + FP + FN)
Radiologist diagnosis on report	88.4 (81.5–93.3)	**100.0** (97.2–100)	81.5 (73.6–87.7)	**100.0** (97.2–100)	89.8 (83.4–94.5)
CNN/Max	81.4 (73.6–87.7)	83.5 (76.2–89.6)	**87.7** (80.6–92.7)	70.8 (61.9–78.2)	85.5 (78.0–90.9)
CNN/Avg	87.6 (80.6–92.7)	95.8 (91.2–98.7)	84.0 (76.2–89.6)	93.7 (88.1–97.3)	89.5 (82.5–93.9)
CNN/Vote	88.4 (81.5–93.3)	97.1 (92.3–99.1)	84.0 (76.2–89.6)	95.8 (91.2–98.7)	90.1 (83.4–94.5)
CNN/LSTM	**89.2** (82.5–93.9)	97.2 (92.3–99.1)	85.2 (78.0–90.9)	95.8 (91.2–98.7)	**90.8** (84.3–95.1)

1024) in our preliminary experiment. In contrast to the feature extraction network that was trained on individual CT slices, the aggregation network was trained on all extracted slices from the middle portion of a CT scan. As discussed in the Feature Aggregation section, we use the confidence score of the last sigmoid layer for the final diagnosis.

Similar to a feature extraction network, this classifier was implemented in PyTorch. The initial learning rate was set to 0.1. CT scans are fed to the network one at a time through the feature extraction network during training, with the aforementioned data augmentation transformations applied on all slices. The dropout technique was also applied in the FC layer with $p = 0.2$. In this training, the FC layers of the feature extraction network were also fine-tuned with the learning rate of 1e-4. After 600 epochs of training, we fine-tuned the network on the combination of training and validation sets with the learning rate of 1e-6 for 50 more epochs. This training took 5 h using an NVIDIA Titan Xp GPU card and a high-performance computer.

Our Rule-based aggregation methods are also implemented within the PyTorch framework to operate on the last layer of the trained feature extraction network as input. As these rule-based methods are deterministic, they do not require additional training or parameter tuning.

3.5. Test set adjudication

To guarantee the quality of our evaluation, the reference standards in our test set are established through two different approaches: a semi-quantitative and a quantitative approach. Both of these approaches are routinely used in clinical studies for assessing vertebral fractures [32]. The assessments in this study were performed by a board-certified radiologist (Y.C.), with over 18 years of musculoskeletal (MSK) radiology experience, who was blinded to the original diagnoses reported in radiology reports and our system's results. In the semiquantitative approach, our domain-expert radiologist reviewed the CT exams in the test set and

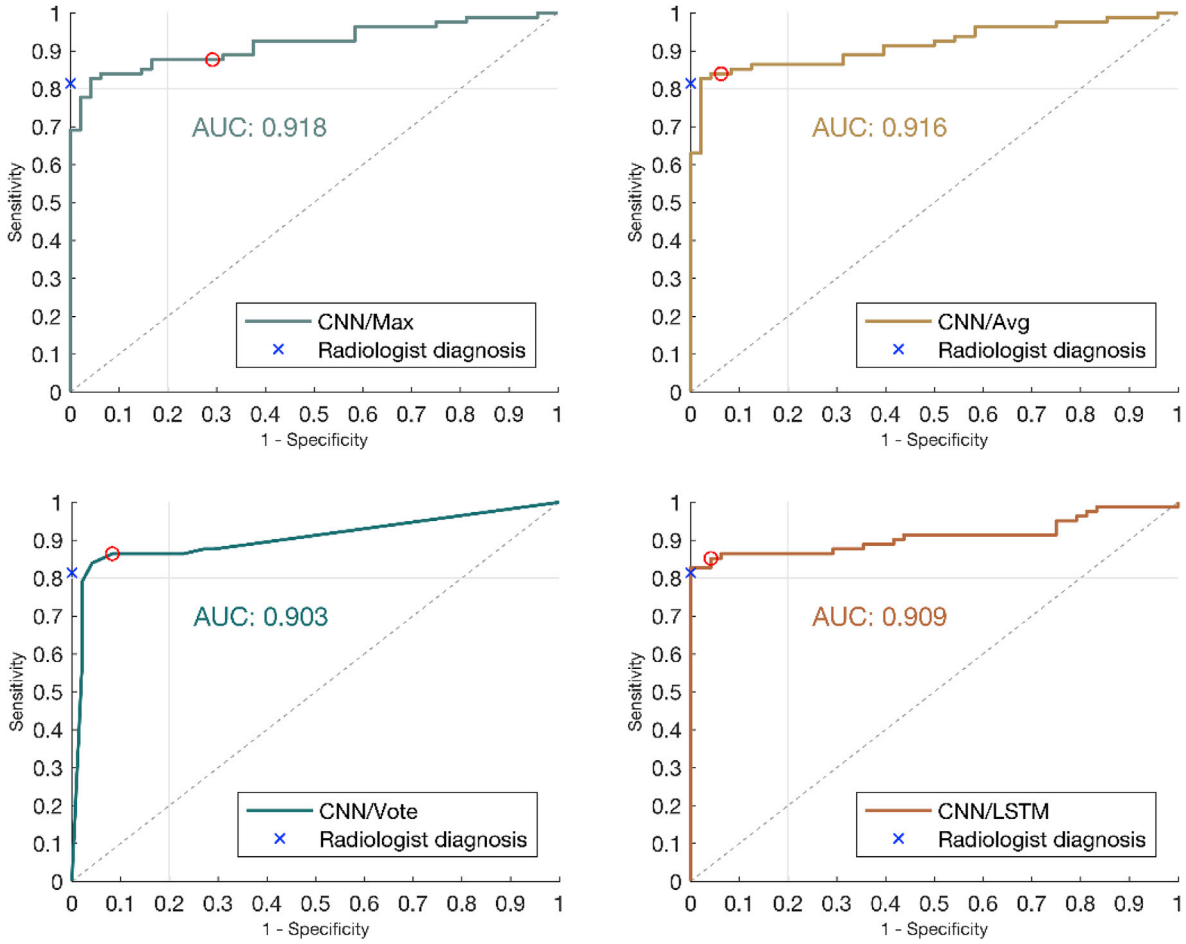

Fig. 3. ROC curves for the different proposed OVF detection systems on our test set. The performance of radiologists on the test set is also marked on the plots based on the corresponding radiology reports. The operating point of the models at 0.5 threshold is marked by a red circle on each curve. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

graded the exams based on a visual inspection. In this well-defined semiquantitative criterion [32], a CT exam is considered positive for OVF, if it is graded 1, 2, or 3.

In cases of disagreements between the semiquantitative assessment and the blinded OVF diagnoses reported in radiology reports or our system's results, additional investigation through a quantitative morphometric approach [32] was performed to establish the reference standard. In this quantitative approach, our domain-expert radiologist measured the vertebral body dimensions using PACS-embedded measuring tools. These measurements were used to calculate the height loss ratio of vertebral bodies. The deformity or wedging with $\geq$20% height loss, which corresponds to grade 1, 2, and 3 in the semiquantitative approach, were considered positive for OVF in our quantitative adjudication. The height loss in our criteria could be anterior, middle, or posterior for a vertebral body. Similar to previous research [4, 9], we observed in this two-phase adjudication, in some cases, that the practicing radiologists did not recognize vertebral fractures on a CT scan. Throughout this adjudication, 16 originally negative cases in our test set were identified as positive for OVF (accounting for 12.4% of the test set).

3.6. Evaluation

Our system outputs a probability value (i.e., a confidence score between 0 and 1) as a diagnosis for each CT scan. A predicted value larger than 0.5 (i.e., middle point probability threshold) is considered positive for OVF on a CT scan, while values less than or equal to 0.5 are considered negative. We evaluated the performance of our system through four standard machine learning metrics on our adjudicated test set: accuracy, precision (positive predictive value), sensitivity (recall), specificity, and F1 score, which is the harmonic mean of precision and recall. The definitions of these evaluation metrics are included in Table 1, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives in our test set.

4. Results

The performance of our proposed models for detecting OVFs on our adjudicated 129 CT scans in the test set is tabulated in Table 1. We also compared the performance of our system to DHMC radiologists' diagnoses as extracted from CT scans' radiology reports in this table. Table 1 shows, among all the proposed models, the CNN/RNN combination achieved the best accuracy and F1 score. Moreover, we observed the accuracy and F1 score of this model matched the DHMC radiologists' performance in detecting OVFs as part of their real-world daily clinical routine.

To further investigate the sensitivity and specificity of our approaches, we plotted the Receiver Operating Characteristic (ROC) curves for our systems in Fig. 3. The ROC curves show our model's performance for different thresholds between 0 and 1. We included the performance of the radiologists on these plots. Although the area under the ROC curve (AUC) was high ($>$0.9) for all approaches, we observed that the ROC for the CNN/LSTM approach overlaps with the radiologists' performance. These results indicate that our CNN/LSTM approach has high efficacy for diagnosing OVF and its performance is on par with practicing radiologists.

We also utilized an occlusion visualization technique [33] to verify the features used by our system to identify OVFs on CT scans were associated with vertebral fractures. Fig. 4 shows four examples of this visualization. The projected color maps on these images highlight the detected fractures on each CT slice.

5. Discussion

In this study, we built a deep neural network model to detect OVFs in chest, abdomen, and pelvis CT exams. The performance of this model matched the performance of practicing radiologists on our test set on multiple evaluation metrics, such as accuracy and F1 score. Our best performing OVF detection approach was composed of a CNN-based neural network to extract features from each CT slice, and an RNN sequence classifier to aggregate these features and make a diagnosis based on the entire CT scan. The visualization of our results and the qualitative investigation of their associated color maps by a domain-expert radiologist showed that these features are on-target for this diagnosis task. Our RNN-based aggregation module is based on the state-of-the-art LSTM architecture. This module is capable of taking into consideration the high-dimensional extracted features from a sequence of CT slices, and, therefore, performs a comprehensive analysis based on the entire CT scan. The comparison of our system to multiple rule-based methods for feature aggregation that relied on slice-level decisions showed that RNN-based aggregation has the highest diagnostic accuracy. We used the performance of the models at their operating point rather

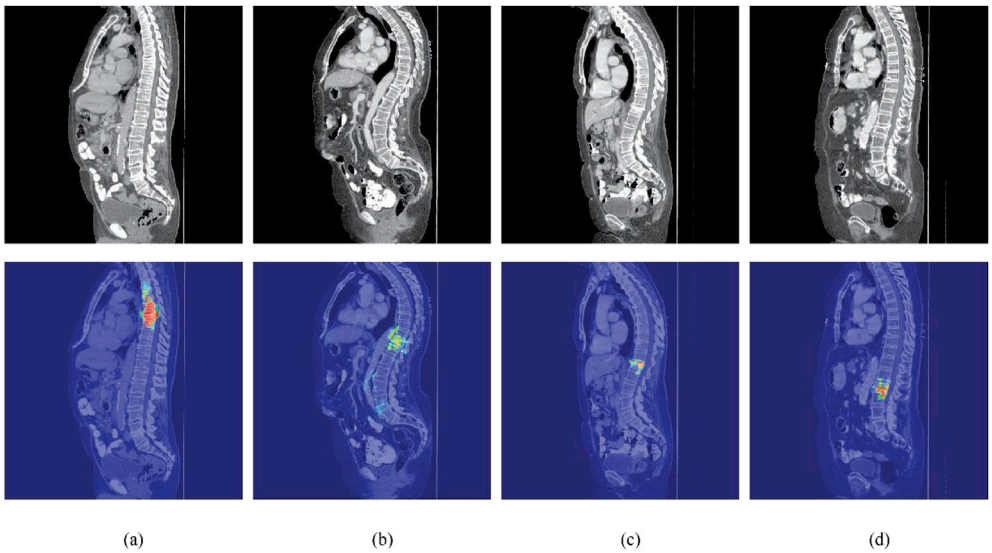

Fig. 4. The top row shows four examples from CTs that were identified positive for OVFs by our CNN/LSTM classification approach with fractures in thoracic vertebrae (a & b) and lumbar vertebrae (c & d). The color maps in the bottom row highlight the regions that are strongly associated with these classification outputs in our model. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

than AUC to compare the model to each other and practicing radiologists. The performance of our CNN/LSTM model showed a well-balanced accuracy and F1 score for the OVF detection task.

Considering the unexpected nature of osteoporotic vertebral fractures on the routine chest, abdomen, and pelvis CT scans, we were extremely careful in our evaluation to avoid introducing any potential biases in measuring a typical radiologist's performance in detecting these fractures. For this purpose, we extracted the diagnoses of the radiologists at the point of care, as documented in the associated radiology reports for CT scans, for the patients in our test set to compare our method to the performance of the radiologists in real clinical circumstances at our institution. Our adjudication established the reference standards for all cases in our test set objectively, through a semiquantitative and a quantitative method. Therefore, in comparison to other common evaluation strategies, in which radiologists are specifically tasked with investigating radiology exams to evaluate a developed detection system, our evaluation strategy produces a more realistic comparison to radiologists' performance in practice at the point of care.

As we observed in our test set adjudication, the discrepancies in radiology report diagnoses are all false negative cases, which are missed by original radiologists. These common missing cases and underreporting can be due to "staff shortages and/or excess workload" [34] or "lack of information about osteoporosis among radiologists" [6]. In addition, OVFs can often be asymptomatic and not be the primary reason for performing CTs [7]. As a result, "busy radiologists tend to pass over incidental findings during dictation in order to save time" [35]. Of note, in our dataset, we cannot rule out the possibility of diagnostic biases introduced by inter-expert variabilities. Although the high performance of our model presented in the Results section shows that the effect of such biases can be limited in our setting, the diagnostic biases and inter-expert variabilities may still affect the performance and generalizability of our approach.

Of note, we used the Chi-square test of statistical significance [36] to compare the performance of our system on the test set to the radiologists' performance. The p-value of this comparison showed that the outperformance of our system over the radiologists is not significant at the 0.05 level of statistical significance. However, our results suggest that our system is at least at the level of practicing radiologists' performance for detecting OVFs.

There has been some previous work on automatic detection of vertebral fractures on CTs using non-deep-learning approaches or combinations of deep learning and non-deep-learning methods [12,19–21]. These methods were comprised of multiple steps on each vertebra on a CT scan to perform the diagnosis. The development of these methods was based on relatively small datasets of CTs (between 15 and 150 CT scans), which were carefully reviewed and selected by domain experts [12,19,20]. These previous methods relied on a non-deep learning segmentation approach to calculate the height loss ratio of each vertebra for the final diagnosis. The resulting systems were fine-tuned on the corresponding development sets, on which the final evaluations were performed as well. As an example, a recent segmentation-based method achieved an accuracy of 88% on a development set of 150 CTs (75 positive and 75 negative cases) for OVF [12]. As another example, a deep learning model that detects OVFs on extracted vertebral bodies along a vertebral column achieved an accuracy of 89.1%, sensitivity of 85.2%, and specificity of 93.8% on a balanced validation set of 250 CTs [21]. While the lack of access to these models and the absence of evaluation on a comparable hold-out test set make a direct comparison between our method and these non-deep learning or hybrid methods difficult, we still observed that our deep learning model tends to achieve a better accuracy on an independent test set of 129 randomly selected CT scans (accuracy = 89.2%, sensitivity = 85.2%, and specificity = 95.8%) in comparison to the best of these methods on their corresponding development sets. To summarize, our method presented in this paper is different from these previous works because (1) the whole system is trained and operates end-to-end over a unified deep neural network framework, and (2) the final diagnosis is made on CT volumes, rather than being based on single CT slices or small slice patches. Using the whole CT volumes for diagnosis in our system is particularly instrumental in cases of patients with deformed spines (e.g., scoliosis). Of note, analysis of each CT scan in the best of the aforementioned models took 5 min on a high-performance computer [12], while our method reduces this time to less than 0.02 s on average for the full analysis of a CT scan, which does not stall the clinical workflow. Therefore, in contrast to the previous methods, our holistic, deep learning approach presents a fast, efficient, and accurate diagnostic tool in this domain.

6. Limitations and future direction

There are several limitations to our work as presented in this paper. Our method was developed and evaluated on data from one academic institution. Thus, further study on data from other institutions is required to show the generalizability of our results, which will be pursued as future work. We also plan to extend our methodology to radiology exams with modalities other than chest, abdomen, and pelvis CTs.

In addition, because in our classification architecture we consider single labels for the entire volume of CT scans, the resulting model is susceptible to learning possible confounding factors in such a classification setting, which may result in diagnostic inaccuracy. As future work, we plan to develop a deep learning-based segmentation architecture to address this drawback. Also, the current architecture does not provide important diagnostic information about the location and grade of the fractures. We expect such segmentation framework will be beneficial for providing this information in future work.

Our current method relies on a heuristic for data extraction that exploits the particular structure of the relevant slices (i.e., the middle 5% of slices) due to human anatomy in order to accelerate training and improve the efficiency of the model (see the Data Preprocessing section). We believe this heuristic can be easily adopted in similar datasets where the relevant information lies in a particular substructure of a sequence of data. As future work, we will explore using spatiotemporal approaches from video action recognition tasks [37–42] to generalize this data extraction step further and to improve the performance of our model. Finally, as future work, we plan to deploy the proposed system in a clinical setting at our institution and conduct a prospective study to evaluate the impact on health outcomes and costs.

7. Conclusions

In this paper, we developed a machine learning approach, fully powered by a deep neural network framework, to automatically detect OVFs on CT scans. The performance of our proposed system on an independent test set matched the level of performance of practicing radiologists at the point of care in both accuracy and F1 score. This automatic detection system can potentially reduce the time and the manual burden on radiologists for OVF screening, as well as reduce the potential false negative errors arising in asymptomatic early stage vertebral fracture diagnoses. We expect the clinical implementation of the proposed technology to result in early detection and treatment of osteoporosis, leading to a decrease in the overall socio-economic burden of osteoporosis, and a significant improvement in associated health outcomes. Moreover, this system will provide a platform to improve the quality of care for rural, small, and poor communities, where access to radiology expertise is limited. Finally, the system presented in this paper can be expanded to other diagnostic analyses in radiology exams, and our team will pursue that in future work.

Ethical considerations

This retrospective single-center machine learning study was designed and conducted according to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [43]. The

use of human subject data in this study was approved by the Dartmouth Institutional Review Board (IRB) with a waiver of informed consent.

Conflicts of interest

None Declared.

Acknowledgment

The authors would like to thank Hoiwan Cheung, Heidi Adams, Joseph Such, and Mellissa Chapman for their help with data collection for this study; Lorenzo Torresani for helpful discussions; and Maksim Bolonkin and Lamar Moss for their feedback on the manuscript.

References

[1] N.C. Wright, A.C. Looker, K.G. Saag, J.R. Curtis, E.S. Delzell, S. Randall, B. Dawson-Hughes, The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine, J. Bone Miner. Res. 29 (2014) 2520–2526, https://doi.org/10.1002/jbmr.2269.

[2] O. Johnell, J.A. Kanis, An estimate of the worldwide prevalence and disability associated with osteoporotic fractures, Osteoporos. Int. 17 (2006) 1726–1733, https://doi.org/10.1007/s00198-006-0172-4.

[3] C. Cooper, T. O'Neill, A. Silman, The epidemiology of vertebral fractures. European vertebral osteoporosis study group, Bone 14 (Suppl 1) (1993), https://doi.org/10.1016/j.jocd.2015.08.004. S89–97.

[4] G.A. Carberry, B.D. Pooler, N. Binkley, T.B. Lauder, R.J. Bruce, P.J. Pickhardt, Unreported vertebral body compression fractures at abdominal multidetector CT, Radiology 268 (2013) 120–126, https://doi.org/10.1148/radiol.13121632.

[5] D. Müller, J.S. Bauer, M. Zeile, E.J. Rummeny, T.M. Link, Significance of sagittal reformations in routine thoracic and abdominal multislice CT studies for detecting osteoporotic fractures and other spine abnormalities, Eur. Radiol. 18 (2008) 1696–1702, https://doi.org/10.1007/s00330-008-0920-2.

[6] T. Bartalena, G. Giannelli, M.F. Rinaldi, E. Rimondi, G. Rinaldi, N. Sverzellati, G. Gavelli, Prevalence of thoracolumbar vertebral fractures on multidetector CT: underreporting by radiologists, Eur. J. Radiol. 69 (2009) 555–559, https://doi.org/10.1016/j.ejrad.2007.11.036.

[7] A.L. Williams, A. Al-Busaidi, P.J. Sparrow, J.E. Adams, R.W. Whitehouse, Under-reporting of osteoporotic vertebral fractures on computed tomography, Eur. J. Radiol. 69 (2009) 179–183, https://doi.org/10.1016/j.ejrad.2007.08.028.

[8] H. Obaid, Z. Husamaldin, R. Bhatt, Underdiagnosis of vertebral collapse on routine multidetector computed tomography scan of the abdomen, Acta Radiol. 49 (2008) 795–800, https://doi.org/10.1080/02841850802165776.

[9] A. Bazzocchi, F. Fuzzi, G. Garzillo, D. Diano, E. Rimondi, B. Merlino, A. Moio, U. Albisinni, G. Battista, G. Guglielmi, Reliability and accuracy of scout CT in the detection of vertebral fractures, Br. J. Radiol. 86 (2013), https://doi.org/10.1259/bjr.20130373.

[10] P.M. Graffy, S.J. Lee, T.J. Ziemlewicz, P.J. Pickhardt, Prevalence of vertebral compression fractures on routine CT scans according to L1 trabecular attenuation: determining relevant thresholds for opportunistic osteoporosis screening., AJR, Am. J. Roentgenol. 209 (2017) 491–496, https://doi.org/10.2214/AJR.17.17853.

[11] S.J. Lee, P.A. Anderson, P.J. Pickhardt, Predicting future hip fractures on routine abdominal CT using opportunistic osteoporosis screening measures: a matched case-control study, Am. J. Roentgenol. 209 (2017) 395–402, https://doi.org/10.2214/AJR.17.17820.

[12] J.E. Burns, J. Yao, R.M. Summers, Vertebral body compression fractures and bone density: automated detection and classification on CT images, Radiology 284 (2017) 788–797, https://doi.org/10.1148/radiol.2017162100.

[13] C.F. Buckens, P.A. De Jong, W.P. Mali, H.J. Verhaar, Y. Van Der Graaf, H.M. Verkooijen, Prevalent vertebral fractures on chest CT: higher risk for future hip fracture, J. Bone Miner. Res. 29 (2014) 392–398, https://doi.org/10.1002/jbmr.2028.

[14] S.J. Lee, P.J. Pickhardt, Opportunistic screening for osteoporosis using body CT scans obtained for other indications: the UW experience, Clin. Rev. Bone Miner. Metabol. 15 (2017) 128–137, https://doi.org/10.1007/s12018-017-9235-7.

[15] P.J. Pickhardt, B.D. Pooler, T. Lauder, A.M. del Rio, R.J. Bruce, N. Binkley, Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications, Ann. Intern. Med. 158 (2013) 588, https://doi.org/10.7326/0003-4819-158-8-201304160-00003.

[16] D.B. Larson, M.C. Chen, M.P. Lungren, S.S. Halabi, N.V. Stence, C.P. Langlotz, Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs, Radiology (2017), https://doi.org/10.1148/radiol.2017170236, 170236.

[17] K. Yasaka, H. Akai, O. Abe, S. Kiryu, Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study, Radiology (2017), https://doi.org/10.1148/radiol.2017170706, 170706.

[18] R. Burge, B. Dawson-Hughes, D.H. Solomon, J.B. Wong, A. King, A. Tosteson, Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025, J. Bone Miner. Res. 22 (2007) 465–475, https://doi.org/10.1359/jbmr.061113.

[19] T. Baum, J.S. Bauer, T. Klinder, M. Dobritz, E.J. Rummeny, P.B. Noël, C. Lorenz, Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT, Eur. Radiol. 24 (2014) 872–880, https://doi.org/10.1007/s00330-013-3089-2.

[20] S. Ghosh, R.S. Alomari, V. Chaudhary, G. Dhillon, in: R.M. Summers, B. van Ginneken (Eds.), Automatic Lumbar Vertebra Segmentation from Clinical CT for Wedge Compression Fracture Diagnosis, 2011, https://doi.org/10.1117/12.878055, 796303.

[21] A. Bar, L. Wolf, O. Bergman Amitai, E. Toledano, E. Elnekave, Compression fractures detection on CT, in: S.G. Armato, N.A. Petrick (Eds.), International Society for Optics and Photonics, 2017, https://doi.org/10.1117/12.2249635, 1013440.

[22] M. Aouache, A. Hussain, M.A. Zulkifley, D.W.M. Wan Zaki, H. Husain, H. Bin Abdul Hamid, Anterior osteoporosis classification in cervical vertebrae using fuzzy decision tree, Multimed. Tool. Appl. 77 (2018) 4011–4045, https://doi.org/10.1007/s11042-017-4468-5.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conf. Comput. Vis. Pattern Recognit, 2016, https://doi.org/10.1109/CVPR.2016.90.

[24] S. Hochreiter, J. Urgen Schmidhuber, LONG SHORT-TERM MEMORY, Neural Comput. 9 (1997) 1735–1780.

[25] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proc. 27th Int. Conf. Mach. Learn, 2010, 10.1.1.165.6419.

[26] V.T. Georgiev, A.N. Karahaliou, S.G. Skiadopoulos, N.S. Arikidis, A.D. Kazantzi, G.S. Panayiotakis, L.I. Costaridou, Quantitative visually lossless compression ratio determination of JPEG2000 in digitized mammograms, J. Digit. Imag. 26 (2013) 427–439, https://doi.org/10.1007/s10278-012-9538-7.

[27] P.Y. Simard, D. Steinkraus, J. Platt, in: Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis, Institute of Electrical and Electronics Engineers, Inc, 2003.

[28] R. Wu, S. Yan, Y. Shan, Q. Dang, G. Sun, Deep Image: Scaling up Image Recognition, 2015.

[29] A. Paszke, S. Gross, S. Chintala, PyTorch, 2017.

[30] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proc. IEEE Int. Conf. Comput. Vis, 2015, pp. 1026–1034.

[31] D.P. Kingma, J. Ba, Adam: a Method for Stochastic Optimization, 2014.

[32] H.K. Genant, C.Y. Wu, C. van Kuijk, M.C. Nevitt, Vertebral fracture assessment using a semiquantitative technique, J. Bone Miner. Res. 8 (1993) 1137–1148, https://doi.org/10.1002/jbmr.5650080915.

[33] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Eur. Conf. Comput. Vis, 2014, pp. 818–833.

[34] A.P. Brady, Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 8 (2017) 171–182, https://doi.org/10.1007/s13244-016-0534-1.

[35] T. Bartalena, M.F. Rinaldi, C. Modolon, L. Braccaioli, N. Sverzellati, G. Rossi, E. Rimondi, M. Busacca, U. Albisinni, D. Resnick, Incidental vertebral compression fractures in imaging studies: lessons not learned by radiologists, World J. Radiol. 2 (2010) 399–404, https://doi.org/10.4329/wjr.v2.i10.399.

[36] F. Yates, Contingency tables involving small numbers and the χ 2 test, Suppl. to J. R. Stat. Soc. 1 (1934) 217, https://doi.org/10.2307/2983604.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proc. IEEE Int. Conf. Comput. Vis, 2015, pp. 4489–4497, https://doi.org/10.1109/ICCV.2015.510.

[38] Y. Wang, M. Long, J. Wang, P.S. Yu, Spatiotemporal pyramid network for video action recognition, in: 2017 IEEE Conf. Comput. Vis. Pattern Recognit, 2017, pp. 2097–2106, https://doi.org/10.1109/CVPR.2017.226.

[39] G.W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2010, https://doi.org/10.1007/978-3-642-15567-3_11.

[40] S. Ji, M. Yang, K. Yu, W. Xu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2013), https://doi.org/10.1109/TPAMI.2012.59.

[41] C. Feichtenhofer, A. Pinz, R.P. Wildes, Spatiotemporal multiplier networks for video action recognition, IEEE Conf. Comput. Vis. Pattern Recognit (2017), https://doi.org/10.1109/CVPR.2017.787.

[42] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential Deep Learning for Human Action Recognition, Springer, Berlin, Heidelberg, 2011, pp. 29–39, https://doi.org/10.1007/978-3-642-25446-8_4.

[43] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, T.B. Ho, S. Venkatesh, M. Berk, Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view, J. Med. Internet Res. 18 (2016), https://doi.org/10.2196/jmir.5870.