Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance

Ghasem Hajianfar [a], Maziar Sabouri [b,c], Yazdan Salimi [a], Mehdi Amini [a], Soroush Bagheri [c], Elnaz Jenabi [d], Sepideh Hekmat [e], Mehdi Maghsudi [c], Zahra Mansouri [a], Maziar Khateri [f], Mohammad Hosein Jamshidi [g], Esmail Jafari [h], Ahmad Bitarafan Rajabi [c], Majid Assadi [h], Mehrdad Oveisi [i], Isaac Shiri [a], Habib Zaidi [a,j,k,l,*]

[a] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva 4, Switzerland
[b] Department of Medical Physics, School of Medicine, Iran University of Medical Science, Tehran, Iran
[c] Rajaie Cardiovascular Medical and Research Center, Iran University of Medical Sciences, Tehran, Iran
[d] Research Center for Nuclear Medicine, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran
[e] Hasheminejad Hospital, Iran University of Medical Sciences, Tehran, Iran
[f] Department of Medical Radiation Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[g] Department of Medical Imaging and Radiation Sciences, School of Allied Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran
[h] The Persian Gulf Nuclear Medicine Research Center, Department of Molecular Imaging and Radionuclide Therapy, Bushehr Medical University Hospital, School of Medicine, Bushehr University of Medical Sciences, Bushehr, Iran
[i] Department of Computer Science, University of British Columbia, Vancouver, BC, Canada
[j] Geneva University Neurocenter, Geneva University, Geneva, Switzerland
[k] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[l] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

Received 28 October 2022; accepted 18 January 2023

Abstract

Purpose: Whole-body bone scintigraphy (WBS) is one of the most widely used modalities in diagnosing malignant bone diseases during the early stages. However, the procedure is time-consuming and requires vigour and experience. Moreover, interpretation of WBS scans in the early stages of the disorders might be challenging because the patterns often reflect normal appearance that is prone to subjective interpretation. To simplify the gruelling, subjective, and prone-to-error task of interpreting WBS scans, we developed deep learning (DL) models to automate two major analyses, namely (i) classification of scans into normal and abnormal and (ii) discrimination between malignant and non-neoplastic bone diseases, and compared their performance with human observers.

Materials and Methods: After applying our exclusion criteria on 7188 patients from three different centers, 3772 and 2248 patients were enrolled for the first and second analyses, respectively. Data were split into two parts, including training and testing, while a fraction of training data were considered for validation. Ten different CNN models were applied to single- and dual-view input (posterior and anterior views) modes to find the optimal model for each analysis. In addition, three different methods, including squeeze-and-excitation (SE), spatial pyramid pooling (SPP), and attention-augmented (AA), were used to aggregate the features for dual-view input models. Model performance was reported through area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, and specificity and was compared with the DeLong test applied to ROC curves. The test dataset was evaluated by three nuclear medicine physicians (NMPs) with different levels of experience to compare the performance of AI and human observers.

* Corresponding author: Habib Zaidi, Ph.D., Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland.
E-mail: habib.zaidi@hcuge.ch (H. Zaidi).

Z Med Phys **xxx** (2023) xxx–xxx
https://doi.org/10.1016/j.zemedi.2023.01.008
www.elsevier.com/locate/zemedi

Results: DenseNet121_AA (DensNet121, with dual-view input aggregated by AA) and InceptionResNetV2_SPP achieved the highest performance (AUC = 0.72) for the first and second analyses, respectively. Moreover, on average, in the first analysis, Inception V3 and InceptionResNetV2 CNN models and dual-view input with AA aggregating method had superior performance. In addition, in the second analysis, DenseNet121 and InceptionResNetV2 as CNN methods and dual-view input with AA aggregating method achieved the best results. Conversely, the performance of AI models was significantly higher than human observers for the first analysis, whereas their performance was comparable in the second analysis, although the AI model assessed the scans in a drastically lower time.

Conclusion: Using the models designed in this study, a positive step can be taken toward improving and optimizing WBS interpretation. By training DL models with larger and more diverse cohorts, AI could potentially be used to assist physicians in the assessment of WBS images.

Keywords: Bone; Whole-body; Scintigraphy; Artificial intelligence; Deep learning

1 Introduction

Whole-body bone scintigraphy (WBS) is used to assess the distribution of methylene diphosphonate (MDP) in bones. Increases or decreases in tracer uptake are caused by various physiological processes occurring in bones, which can be indicators of malignant or benign diseases under some circumstances. Overall, clinical indications for WBS can be divided into three categories: (i) when a specific bone disorder exists or is suspected, (ii) handling of unexplained symptoms, and (iii) metabolic evaluation before the start of the treatment [1,2]. Despite the low specificity, various advantages, such as high sensitivity, low cost, and the ability to image the whole body effortlessly, make WBS the most suitable imaging modality for studying oncologic patients with skeletal disease involvement [3].

Non-neoplastic diseases include a series of disorders that involve the entire skeleton. These disorders generally increased absorption of bisphosphonate and bone turnover [4–7]. In the early stages of these disorders, it might be challenging to interpret WBS scans as they often reflect normal appearance prone to subjective interpretation. As a result, WBS is seldom used to detect non-neoplastic diseases in the early stages [5–7]. Malignant bone diseases include bone metastasis and rare primary bone tumours. Bone metastasis commonly occurs in patients with frequent solid tumours, such as lung, breast, prostate, and thyroid [8–12], with 65% of them having roots in breast and prostate cancers for women and men, respectively [13]. Bone metastasis mainly affects the spine, femur, and pelvis, though it is not limited to these regions [14]. It also adversely affects several characteristics of the patients, such as survival, morbidity, and quality of life. In addition, it can provoke skeletal complications, such as skeletal remodelling, pathologic fractures, pain, and anemia [11]. Therefore, early and accurate diagnosis of malignant disease is essential in survival improvement and treatment management [12,15,16].

Several imaging modalities are currently available for the detection of malignant bone diseases, including bone scintigraphy, radiography, magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET)/CT [17–19]. Among the various strategies, WBS is the most common method for detecting bone involvement, owing to its high sensitivity (95%) and ability to screen the whole body in one session [1]. However, despite its high sensitivity, WBS does not provide high specificity in the sense that it is less reliable in distinguishing malignant disorders from other causes of increased bone turnover, such as osteomyelitis, healing fracture, etc. [20]. Moreover, interpreting WBS scans is time-consuming, increases workload, and requires considerable experience [14]. Hence, automated methods for clinical diagnosis and classification of bone diseases using WBS scans are highly significant.

Recently, deep learning (DL) algorithms driven by developments in artificial intelligence (AI) have shown promising results toward medical images analysis [21–26]. In particular, convolutional neural networks (CNNs) showed great potential in analysing and classifying medical images [14,17,18,27–29]. Although there has not been much research on using DL to analyse WBS, few studies have reported on the promising potential of DL algorithms in diagnosing malignant bone diseases from WBS images [14,17,27,28,30–33]. For example, Liu et al. [33] developed a DL-based method to automatically evaluate bone metastases on bone scintigraphy. They concluded that the accurate identification and automatic analysis of bone metastases is possible using DL. Han et al. [31] also investigated the performance of DL to classify bone scan images in patients with prostate cancer, demonstrating excellent discriminative performance (presence vs. absence of metastases). Papandrianos et al. [14,17,27] developed DL methods for diagnosing

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

malignant bone diseases using WBS scans in breast and prostate cancer patients, reporting a promising performance. Their results showed that the DL model accurately distinguishes malignant bone diseases from degenerative changes and normal tissue [14,17,27]. Zhao et al. [28] claimed that their AI model saves time and improves diagnostic accuracy for malignant bone diseases. In the study by Pi et al. [18], high classification accuracy was achieved, indicating the effectiveness of their proposed architecture for interpreting WBS, suggesting that their model can be used as a clinical decision support tool. Hence, DL can be utilized to conquer WBS scanning challenges, such as the need for experience and time for image interpretation and reporting.

WBS produces two planar scans, including anterior and posterior views. These two views may contain complementary information for the diagnosis of bone complications. Thus, an optimum model must simultaneously analyse both views to exploit comprehensive information. Recently, promising results have been obtained by DL methods with multi-view inputs, specifically in applications relating to mammography image analysis [34–36]. Concerning clinical applications, Wang et al. [37] and Liu et al. [38] developed multi-view CNN models using axial, sagittal, and coronal views of CT images for lung nodule segmentation and classification, respectively. Inspired by these studies, Pi et al. [18] and Zhao et al. [28] designed a methodology toward automated diagnosis of malignant bone diseases using multi-view strategies.

In this work, we developed multiple DL models to achieve optimum performance targeting two main applications of WBS scanning using different CNN algorithms and multi-view aggregation methods. The first part focused on classifying patients into normal and abnormal subjects, whereas the second part focused on discriminating patients with malignant bone diseases from patients diagnosed with other abnormalities (non-neoplastic). We compared the performance of DL models with human observers for these two tasks. The models proposed in this study can be used to reduce the burden of WBS scan interpretation in clinical setting.

2 Materials and methods

2.1 Patients selection

In this retrospective multicenter study, 7188 patients referred to WBS with ^{99m}Tc-MDP from 1 October 2015 to 30 September 2019 were enrolled. However, patients with incomplete or inaccessible records, low-quality images, and those who had subcutaneous injections and skin surface contamination were excluded. After applying the inclusion and exclusion criteria, shown as a flowchart in Fig. 1,

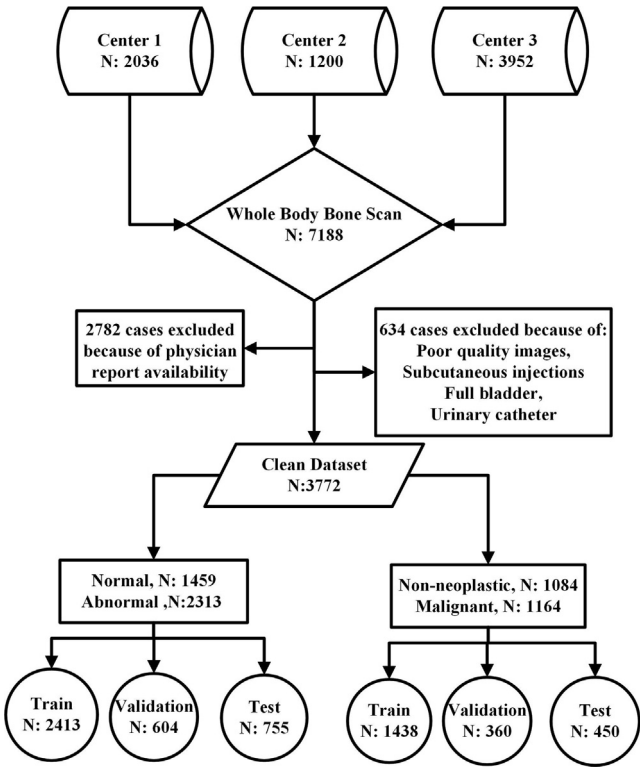

Figure 1. Flowchart of inclusion and exclusion criteria.

3772 patients were enrolled in the first analysis. In the second analysis, from the 2313 abnormal cases, 65 cases were removed because they did not have a definite report to distinguish abnormality. Accordingly, the remaining 2248 abnormal cases were used in the second analysis. Table 1 shows the baseline characteristics of the enrolled patients. Nuclear medicine physicians' reports based on patient's history, lab tests, pathology, and current WBS with additional spot images were considered as ground truth conducted by two nuclear medicine physicians' readers. Abnormal cases were patients who had one of the non-neoplastic bone diseases (degenerative, infectious, spondylosis, traumatic lesions, and inflammation) or malignant bone diseases (bone metastasis and primary bone tumour); nevertheless, normal cases did not belong to the abnormalities mentioned. Fig. 2 shows cases for each category of normal and other disorders included in this study. Two different analysis strategies were implemented in this study. The purpose of the first one was to distinguish normal from abnormal cases based on physicians' reports. The abnormal category included non-neoplastic bone diseases and malignant bone diseases. The second analysis intended to diagnose non-neoplastic disorders against malignant bone diseases. Fig. 1 shows the partitioning of patients for each analysis.

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

Table 1
Characteristic of patients enrolled in the study protocol from the different clinical centers.

Characteristic	Center 1	Center 2	Center 3	Total
Number	1152	701	1919	3772
Female	730 (63%)	450 (64%)	1302 (68%)	2482 (66%)
Male	422 (37%)	251 (36%)	617 (32%)	1290 (34%)
Age (Mean ± Sd)	52.83 ± 17.53	52.51 ± 17.72	50.56 ± 17.21	51.61 ± 17.40
Age Range	1.5-95	1-90	1.5-88	1-95
Bone Status				
Normal	269	216	976	1461
Degenerative	237	138	196	571
Inflammation	96	56	53	205
Osteomyelitis	6	5	24	35
Trauma	136	70	92	298
Tumor	38	22	91	151
Metastatic	370	194	487	1051

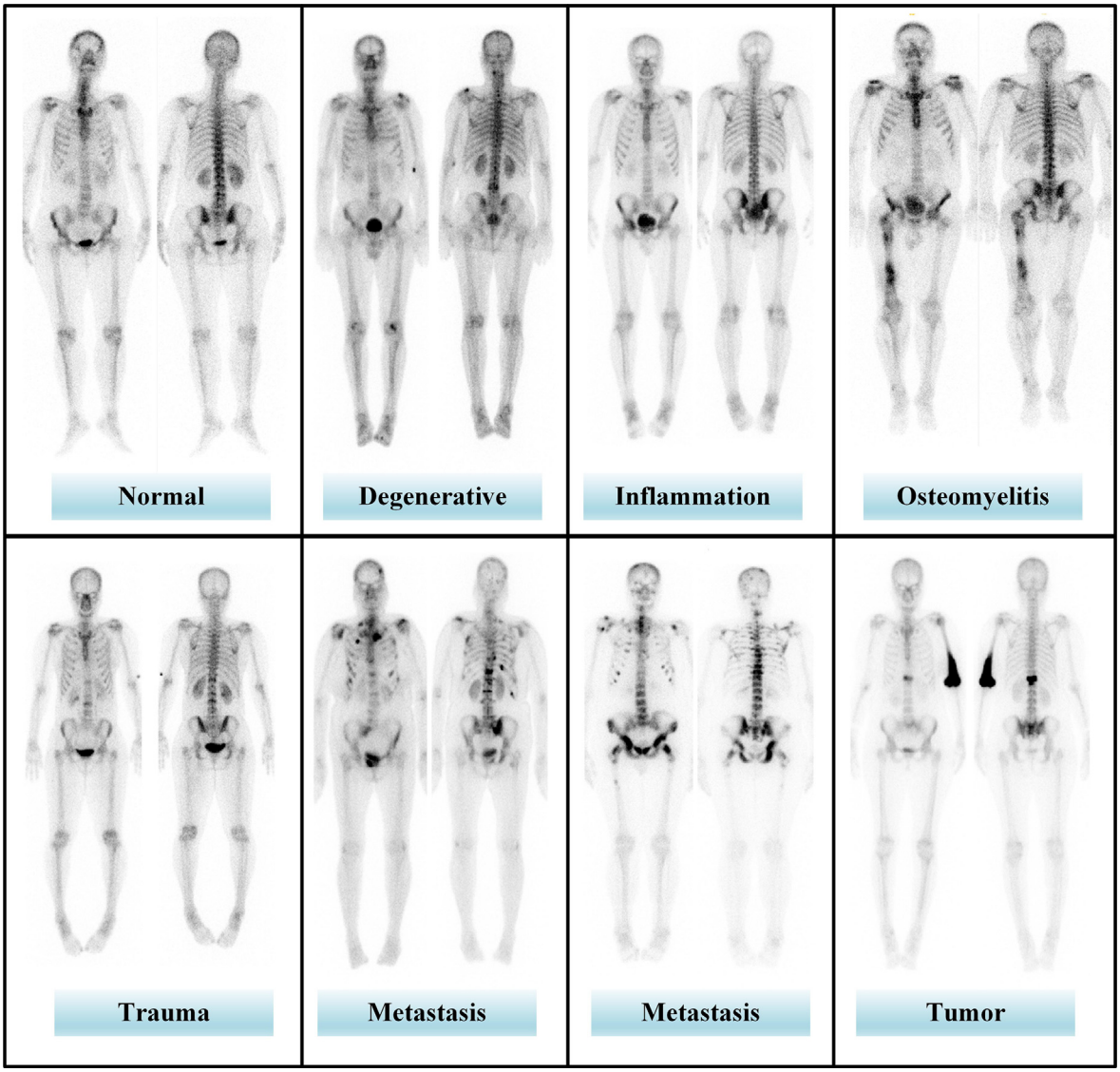

Figure 2. An instance of normal and pathological cases according to nuclear medicine physicians' reports.

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

2.2 Bone scintigraphy imaging

WBS scans were performed 2-4 hours post-injection following intravenous injection of 555 to 925 MBq of ^{99m}Tc-MDP on a dual-head gamma camera (Siemens Symbia Encore, Siemens ECAM IP1 and Mediso AnyScan S) equipped with low-energy parallel-hole high-resolution collimators in supine arm-down position. The energy acquisition window was centered on 140 KeV with a 20% window, a scan velocity of 12-15 cm/min in continuous mode, and a matrix size of 1024 × 256.

2.3 Deep learning workflow

The data were split into three subsets, including train (60%), validation (20%), and test (20%). Initially, all images in the training dataset were resized to 768 × 256, followed by normalization according to the maximum intensity. Table 2 shows the number of cases used as train, validation, and test sets for each analysis.

For each analysis, two strategies were pursued: single input (anterior or posterior) and dual input. For the single input strategy, ten CNN models including VGG19, MobileNetV2, ResNet50V2, ResNet101V2, ResNet152V2, InceptionV3, InceptionResNetV2, DenseNet121, DenseNet169 and DenseNet201 with squeeze-and-excitation (SE) [39] were trained. For the dual input strategy, the combination of the 10 aforementioned CNN models with three aggregation methods (dual input), including SE, spatial pyramid pooling (SPP) [40], and attention-augmented (AA) [41], was utilized. Hence, for each analysis, 50 different models containing 20 from single-input strategy (10 models with anterior and 10 models with posterior views) and 30 from dual-input strategy (cross-combination of 10 CNN and 3 aggregation methods) were developed. Fig. 3 shows the DL workflow. A fully connected layer for SPP and SE inspired by [18,28] was utilized. All models were developed in Python 3.6 (TensorFlow 2.2 and Keras 2.4.3) and Linux operating system (Ubuntu 18.04) workstation with NVIDIA GEFORCE 1080Ti with 11 GB of RAM.

2.4 Evaluation of the DL models performance

The results of the models were evaluated by four parameters, including accuracy (ACC), area under receiver operating characteristic (ROC) curve (AUC), sensitivity (SEN), and specificity (SPE). We also compared the performance of the proposed AI methods with humans' performance. The same analysis performed with DL methods was also performed by three nuclear medicine physicians (NMPs) with different levels of experience who were blind to the results. This includes a 4th year resident (NMP1), one with 5 years of experience (NMP2), and one with 21 years of experience (NMP3). They only had access to the anterior and posterior WBS images without access to patients' history or other examinations/scans. The anonymized images in DICOM format were transferred to a medical imaging workstation in one of the nuclear medicine imaging centers and were viewed on a dedicated monitor. The workstation allowed physicians to change the brightness, window level, gamma, and zooming while offering different look-up tables (color maps). In addition, they were allowed to perform quantitative measurements, e.g., line profiles and regions of interest analysis. Comparisons between AUCs achieved by DL methods and NMPs were performed by the DeLong test followed by false discoveries rate (FDR) correction with the Benjamini Hochberg method applied on p-values, and adjusted p-values (q-values) were reported. The p-values (or q-values) less than 0.05 were considered statistically significant.

Table 2
Distribution of data in two analysis and evaluation strategies.

Dataset	Total	Train	Validation	Test
Analysis #1	3772	2413	604	755
Normal	1459	928	254	277
Abnormal	2313	1458	350	478
Analysis #2	2248	1438	360	450
Non-neoplastic	1084	692	171	221
Malignant	1164	746	189	229

3 Results

3.1 Normal vs. abnormal classification

Fig. 4 shows the results of various DL models used for the first analysis. The best 5 models achieving the highest performance (considering a compromise between performance metrics) were identified. The accuracy, AUC, sensitivity, and specificity achieved by the different models were 0.66, 0.69, 0.58, and 0.80 for the InceptionV3_Ant model, 0.66, 0.67, 0.63, and 0.71 for ResNet50V2_Post model, 0.72, 0.72, 0.73, and 0.70 for DenseNet121_AA model, 0.72, 0.68, 0.82, and 0.54 for InceptionResNetV2_SE model and finally 0.70, 0.70, 0.69, and 0.72 for InceptionV3_ SPP model. The ROC curves of the best 5 models for the first analysis are illustrated in Fig. 5-a-e. Fig. 5f compares ROCs achieved by NMPs and DenseNet121_AA, which achieved the highest AUC among all DL models.

3.2 Malignant vs non-neoplastic discrimination

Fig. 6 shows the results of various models used in the second analysis. The accuracy, AUC, sensitivity, and specificity achieved by the 5 best performing models were 0.67, 0.67, 0.59, and 0.76 for DenseNet201_Ant, 0.64, 0.64, 0.62, and 0.66 for DenseNet121_Post, 0.70, 0.70, 0.61, and 0.79 for

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

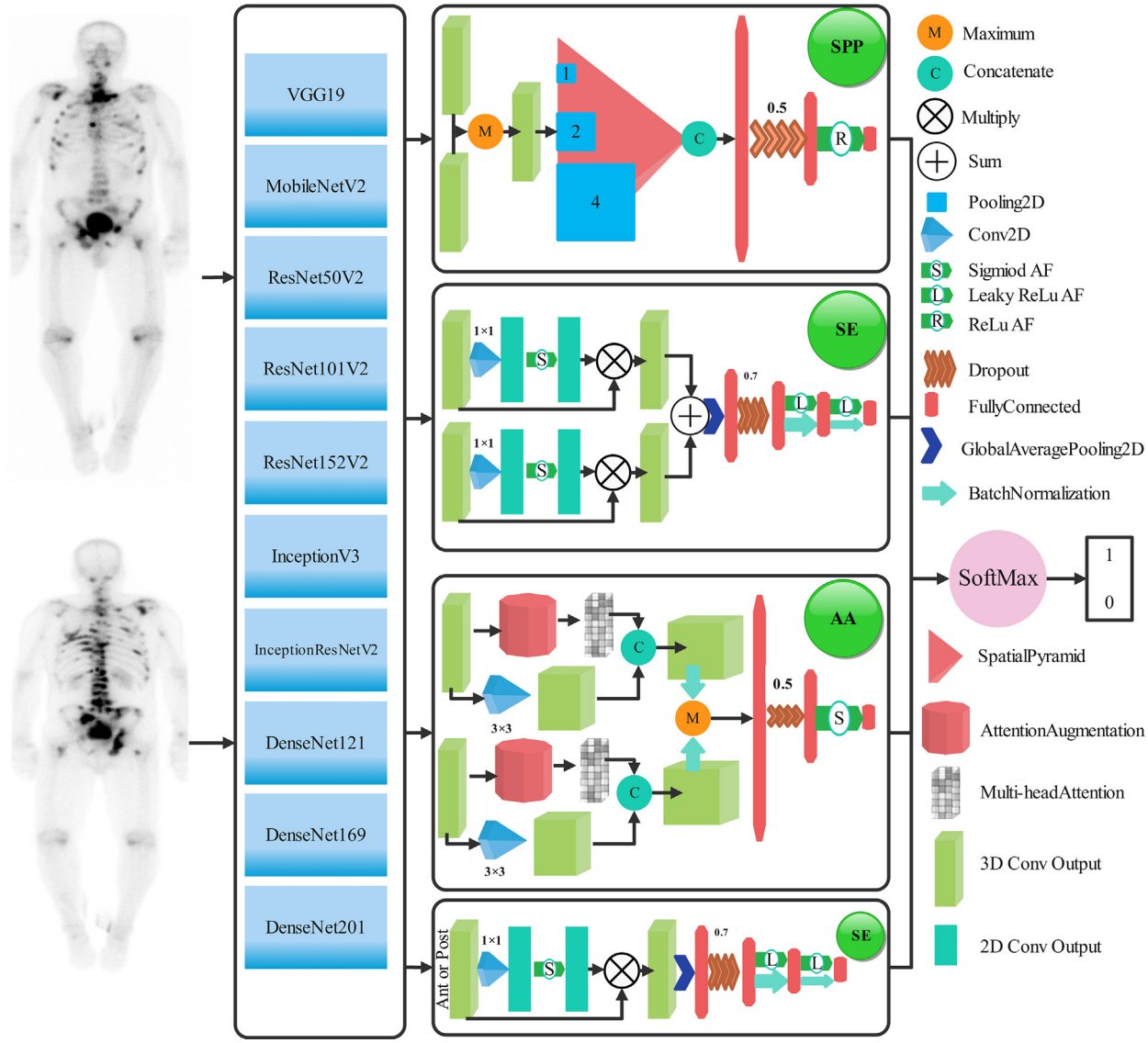

Figure 3. Workflow of applied deep learning models. Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

InceptionResNetV2_AA, 0.64, 0.64, 0.52, and 0.76 for DenseNet121_SE, 0.71, 0.72, 0.55, and 0.88 for InceptionResNetV2_SPP model. ROC curves of the best 5 models for the second analysis are illustrated in Fig. 7a-e. Fig. 7f compares the ROC curves achieved by NMPs and InceptionResNetV2_SPP, which achieved the highest AUC among all DL models.

Since several models with various modes were employed in this study, the best performing model has to be identified. For this purpose, we did not limit ourselves to reporting conventional metrics, such as AUC, ACC, SEN, and SPE. Instead, the best performing models with respect to these four metrics were also assessed in terms of the DeLong test. All the models were compared using the DeLong test, run on the model's AUC, to identify the best models with a higher margin of confidence. Therefore, Densenet121-AA of the first strategy and InceptionResNetV2-SPP of the second strategy were determined to be the best models.

For both analyses, the DeLong test was utilized to compare the performance of each DL model with the remaining 49 models (a total of 50 DL models were developed for each analysis). The results of these comparisons are presented as binary significant/non-significant (color-coded) in Figs. 8 and 9 for the first and second analyses, respectively. In addition, for both strategies, the results of DeLong's top five models are presented in tabular format in Tables 3 and 4.

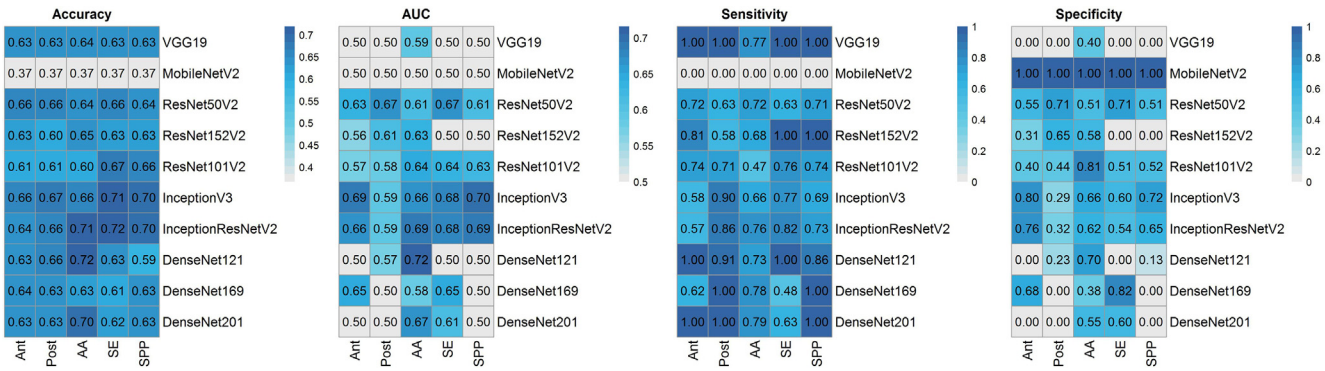

Figure 4. Performance of the various models used in the first analysis in terms of accuracy, AUC, sensitivity, and specificity. Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

Within the first analysis, the performance of Densenet121-AA (achieved the highest AUC in the first analysis) was significantly higher than 43 models. However, it had comparable results with InseptionV3-SPP, InseptionV3-SE, InseptionV3-Ant, InceptionResNetV2-SPP, InceptionResNetV2-SE, and InceptionResNetV2-AA models.

In the second analysis, the performance of InceptionResNetV2-SPP (achieved the highest AUC in the second analysis) was significantly higher than 46 models. However, it performed comparably with InseptionV3-SPP, InceptionResNetV2-AA, and DenseNet201-Ant models.

Table 5 summarizes the average performance (in terms of AUC) achieved by AI models for various CNN architectures and inputs (anterior, posterior, AA, SE, and SPP) for each analysis strategy. For example, in analysis one, Inception V3 and InceptionResNetV2 CNN models and dual-view input with AA aggregating method achieved superior performance on average. In the second analysis, DenseNet121 and InceptionResNetV2 as CNN methods and dual-view input with AA aggregating method achieved the best performance.

3.3 Comparison of DL against human performance

To compare the performance of the proposed DL models against human observers for the first analysis, 755 patients from the test dataset were presented to three NMPs for evaluation. NMP1/NMP2/NMP3 achieved accuracy, AUC, sensitivity, and specificity of 0.72/0.72/0.68, 0.65/0.64/0.60, 0.72/0.93/0.69, and 0.70/0.35/0.65, respectively. Conversely, the DenseNet121_AA model achieved 0.72, 0.72, 0.73, and 0.70, respectively (Table 6). In addition, the DeLong test was performed to compare the performance of DenseNet121_AA and three NMPs with respect to the ROC metric. DenseNet121_AA significantly outperformed all NMPs ($p < 0.05$) (Table 7). In addition, the performance of NMP1 vs. NMP3 (p-value < 0.005) and NMP2 vs. NMP3 (p-value < 0.01) were significantly different.

Regarding the second analysis, 450 patients from the test dataset were used to compare AI and human observers. NMP1/NMP2/NMP3 achieved accuracy, AUC, sensitivity, and specificity of 0.74/0.70/0.77, 0.74/0.70/0.77, 0.85/0.79/0.83, and 0.68/0.62/0.72, respectively. Conversely, InceptionResNetV2_SPP achieved 0.71, 0.72, 0.55, and 0.88 for accuracy, AUC, sensitivity, and specificity, respectively (Table 6). Similar to the first analysis, the DeLong test was performed to compare the performance of three NMPs and AI models with respect to the ROC metric. Unlike the first analysis, except for NMP2 vs. NMP3 (p-value = 0.026), the performance of neither of them was significantly different from the others (Table 7). In addition, the comparison of true classifications within the different types of abnormalities is shown in Table 8 for NMPs and AI models. In the first strategy, there were 214, 213, and 242 cases misdiagnosed by NMP1, NMP2, and NMP3, respectively (false negative and false positive). Nevertheless, our model correctly diagnosed 125, 139, and 154 of the wrongly diagnosed cases by NMPs. In addition, 212 patients were wrongly diagnosed by our model, of which NMP1 NMP2 correctly diagnosed 159, 176 and 170, and NMP3, respectively. In the second strategy, 117, 133, and 105 cases were misdiagnosed by NMP1, NMP2, and NMP3, respectively. However, AI could correctly diagnose 30, 26, and 25 of these cases. Moreover, 129 patients were wrongly diagnosed by our model, of which 39, 80, and 55 were correctly diagnosed by NMP1, NMP2, and NMP3, respectively.

4 Discussion

Although WBS remains the most suitable imaging modality for clinical diagnosis of malignant bone diseases during the early stages, the procedure inherently bears a number of challenges. The procedure is time-consuming and requires vigour and experience [12,14–16]. Moreover, interpretation of WBS scans in the early stages of the disor-

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

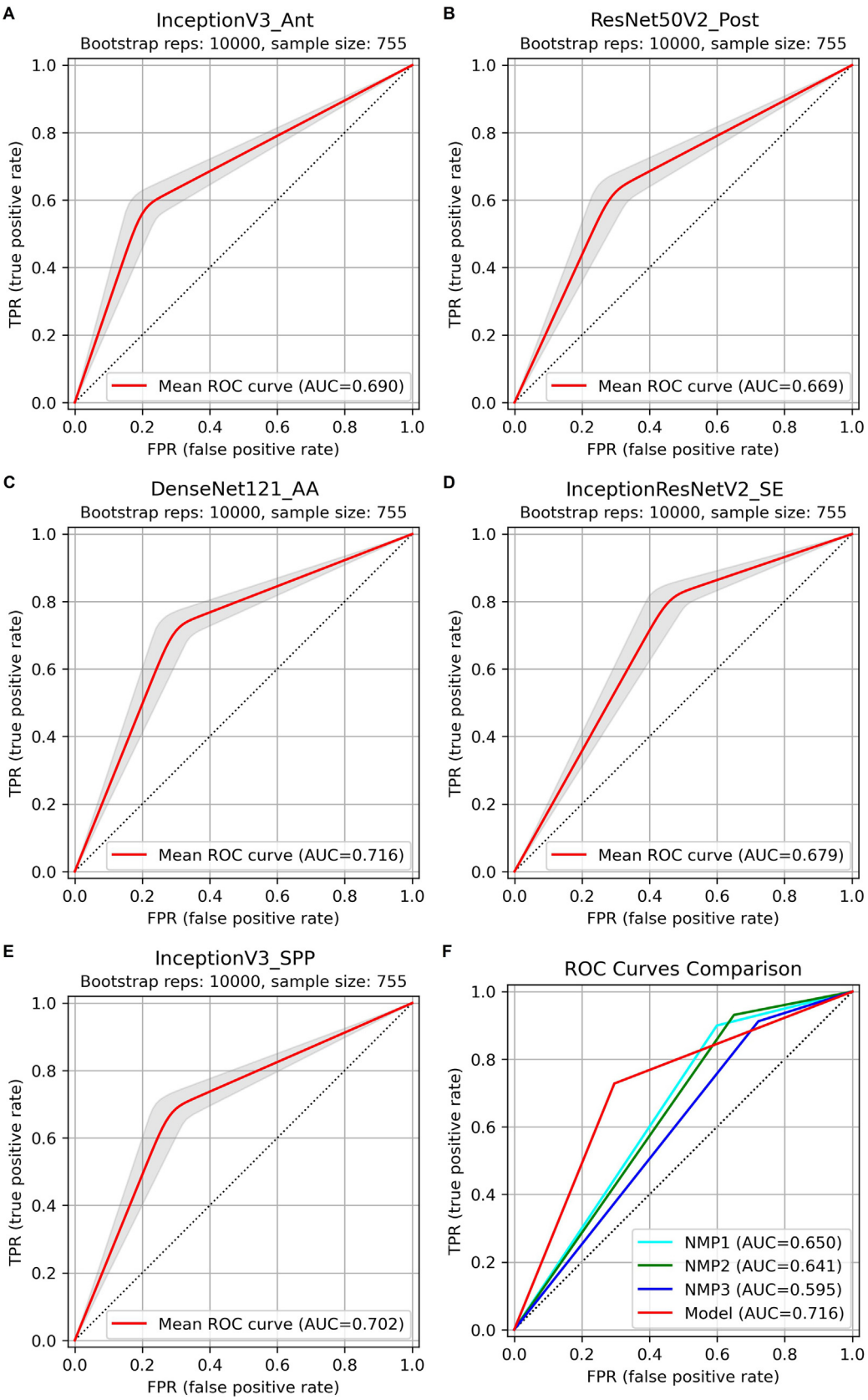

Figure 5. (a-e) ROC curves of the best 5 models for the first analysis. (f) Comparison between the ROC curves achieved by nuclear medicine physicians (NMPs) and the DL model achieving the highest AUC (DenseNet121_AA). Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

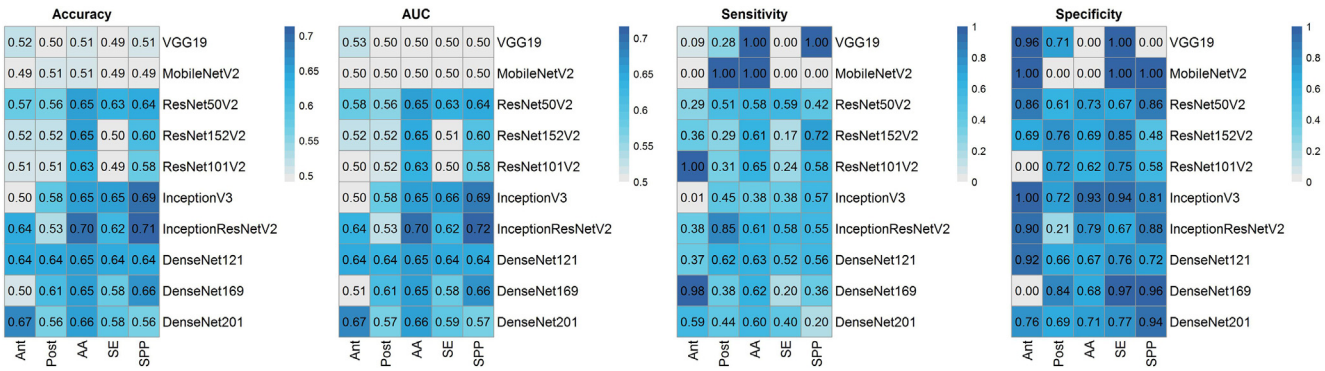

Figure 6. Performance of the various models used in the second analysis regarding the accuracy, AUC, sensitivity, and specificity. Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

ders may be challenging because they often reflect a normal appearance which is subject to subjective qualitative interpretation [5–7]. As such, the development of automated models that can be exploited by nuclear medicine physicians to interpret WBS images is highly demanded. Innovations in computer-aided diagnostic models to analyse WBS images have been pursued for decades. Studies in the literature can be categorised into three main branches [18], including automatic bone scan index (BSI) calculation [42], delineation of bone lesions [43–45], and automated diagnosis of bone metastasis [46,47]. However, compared to traditional image processing approaches for assisting the interpretation of WBS scans, DL is considered more efficient and robust, owing to their independence from segmentation of regions of interest and automatic extraction of image features rather than using hand-crafted ones. Nevertheless, few studies have used deep neural networks to develop automated models for the diagnosis of malignant bone diseases from WBS images.

This study was orchestrated to address the above-mentioned challenges comprehensively. We pursued two main analysis strategies, including the discrimination between normal and abnormal WBSs and patients with malignant bone diseases from non-neoplastic disorders. In the first analysis, the purpose was to design a model enabling to decrease the time and workload spent by physicians for the discrimination between normal and abnormal patients. However, our vision is that such models will be able to classify all bone lesions independently in the near future. Therefore, the intention behind conducting the second analysis was to address a challenge faced in clinical nuclear medicine: the differentiation of malignant bone diseases from non-neoplastic lesions. Two different modes were considered to find the optimal model for each analysis strategy: single-view input (anterior or posterior) and dual-view input. In addition, three different aggregating methods were utilized for the dual-view input mode. Finally, 10 different CNN models were developed for each mode resulting in 50 different models for each analysis.

In a study by Pi et al. [18], three DL models, including Inception-V3, DenseNet-169, and SE-ResNet-50, were used along with SE feature aggregation to detect malignant bone diseases from WBS scans. They enrolled 16,211 patients and used both single- and multi-view inputs using the Inception-V3 model. Their best performance was achieved by Inception-V3 combined with the SE method to aggregate multi-view inputs. They reported accuracy, sensitivity, and specificity of 95%, 93.17%, and 96.1%, respectively. In the study of Zhao et al. [28], 12,222 patients were enrolled where the ResNet-50 model with SPPs was used to detect malignant bone diseases using WBS. They reported AUC for the diagnosis of different types of cancer of 95.5%, 98.8%, 95.7%, and 97.1% for prostate, breast, lung, and other, respectively, reaching an overall AUC of 96.4%. In the study by Hsieh et al. [32], 19,041 patients were used to implement and evaluate the DL framework. Each bone scan image consisted of anterior and posterior planar images used simultaneously for each patient. The posterior and anterior images were first merged and fed to the models. They used CNN, DenseNet121, and ResNet50V2 models with and without Supervised Contrastive Learning (SCL) to detect bone metastasis. Their best model in both conditions (with and without SCL) was ResNet50V2 (ACC = 0.957, SEN = 0.533, SPE = 0.995 and ACC = 0.961, SEN = 0.599, SPE = 0.993, respectively). In their study, Aoki et al. [30] compared the performance of DL for detecting bone metastases in prostate cancer patients with nuclear medicine specialists. In this study, 139 prostate cancer patients were evaluated. At first, a DL-based program was used to segment and extract hot spots. The program used butterfly-type networks (Btrfly-Nets) to fuse anterior and posterior images. The nuclear medicine specialists achieved a sensitivity, specificity, and accuracy of 100% (60 of 60),

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

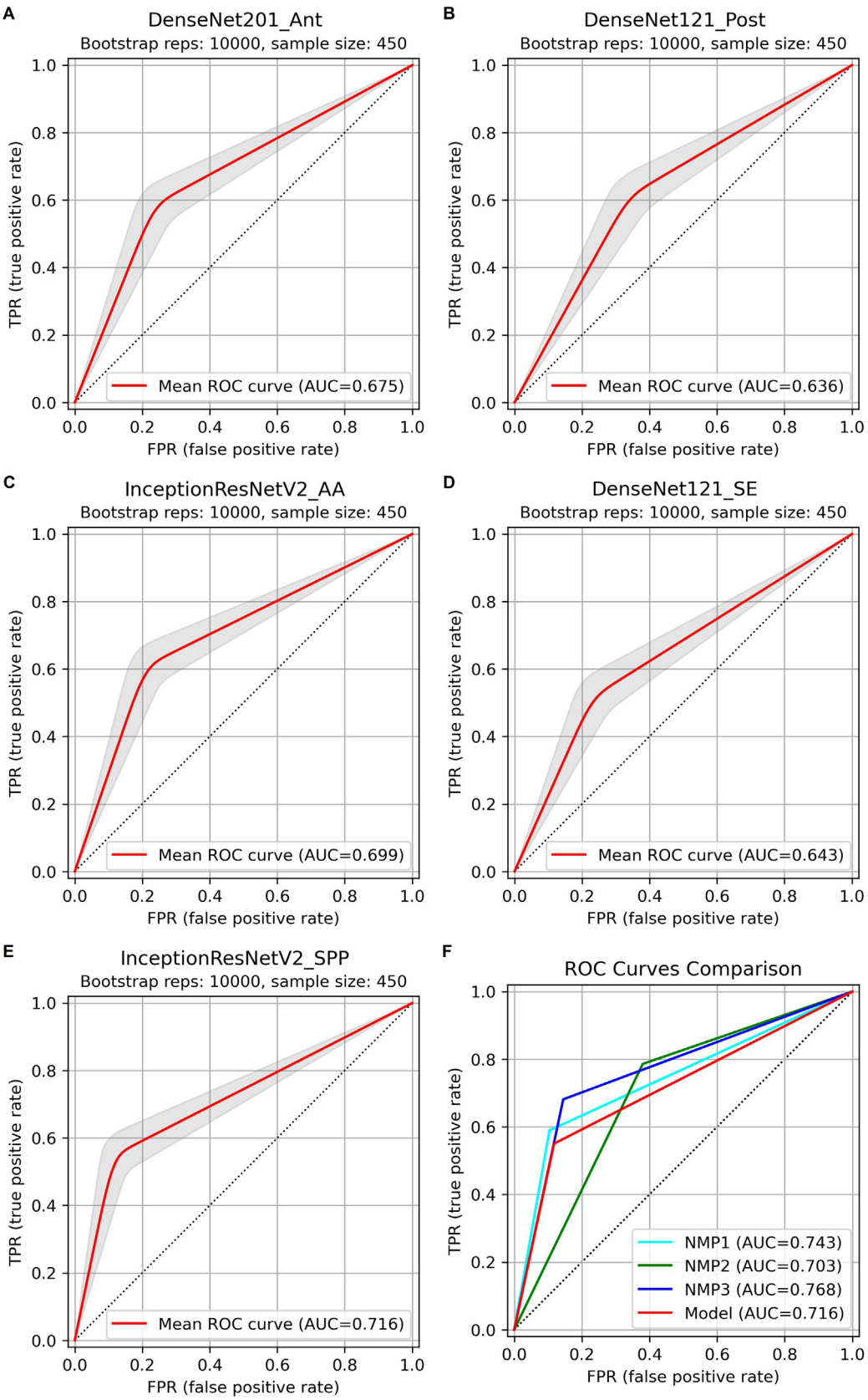

Figure 7. (a-e) ROC curves of the best 5 models for the second analysis. (f) Comparison between the ROC curves achieved by nuclear medicine physicians (NMPs) and the best-performing DL model (InceptionResNetV2_SPP). Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

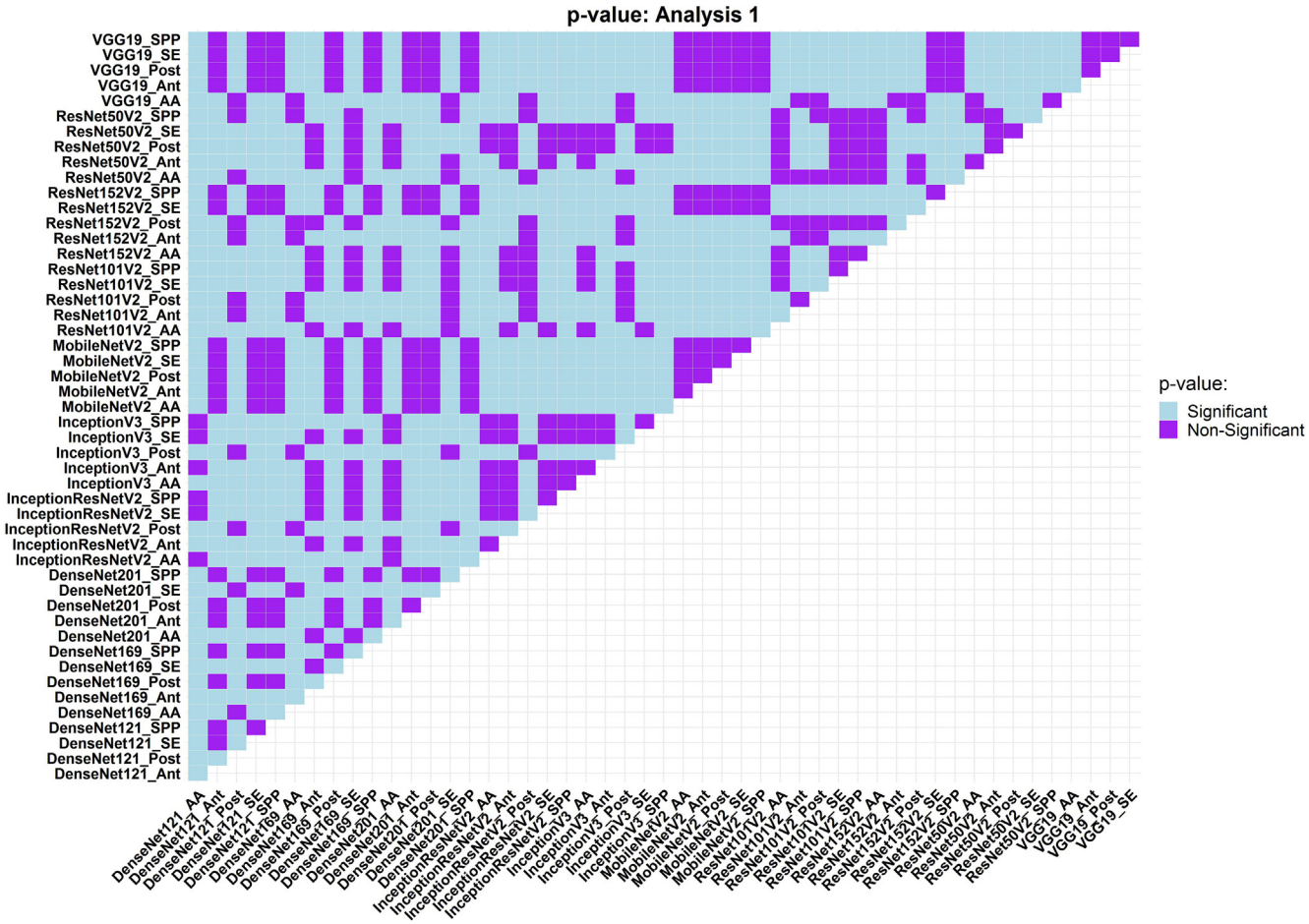

Figure 8. Model performance is compared using the DeLong test for the first strategy, which is run on the models' AUCs. The models on columns and rows were evaluated against each other. Light blue: if the row model outperformed significantly the column model in terms of p-value. Purple: if the comparison between the row model and column model yielded a non-significant p-value. Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

94.9% (75 of 79) and 97.1% (135 of 139), respectively, while the DL approach reached 91.7% (55 of 60), 87.3% (69 of 79) and 89.2% (124 of 139), respectively.

A DL architecture for automatic bone metastases interpretation was suggested by Liu et al. [33] to classify, segment, and extract features of metastatic bone lesions from bone scintigrams and create preliminary reports automatically. ResNet34 was used for classification, whereas U-net encoding and decoding were applied to speed up model's convergence and segmentation. The performance of their model was compared with 3 NMPs with different experience (less than 2 years, 5 years, and 10 years of experience). The AUC of the model was 0.9263, whereas the accuracy was 88.62% which had no meaningful difference with NMPs with 5 and 10 years of experience. However, it was significantly better than the less experienced NMP. Han et al. [31] used 9133 bone scans encompassing 2991 patients with bone metastases and 6142 without, to feed 2D-CNN frameworks, including whole body–based (WB) and tandem frameworks using both whole-body and local patches, called global–local unified emphasis (GLUE). The data were divided into train, validation, and test using two strategies, namely 72%, 8%, and 20% for training, validation, and test sets (abundant training data set) and 10%, 40%, and 50% (limited training dataset). Using abundant training dataset, the AUC of WB and GLUE models were comparable (GLUE: 0.936–0.955, WB: 0.933–0.957, p-value > 0.05 in 4 of 5-fold). When using limited training dataset, the GLUE model outperformed WB (GLUE: 0.894–0.908, WB: 0.870–0.877, p-value < 0.0001).

In comparison, our best model for the diagnosis of malignant disease was InceptionResNetV2 linked with SPP for input combination, which achieved an AUC, accuracy, sensitivity, and specificity of 72%, 71%, 55%, and 88%, respec-

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

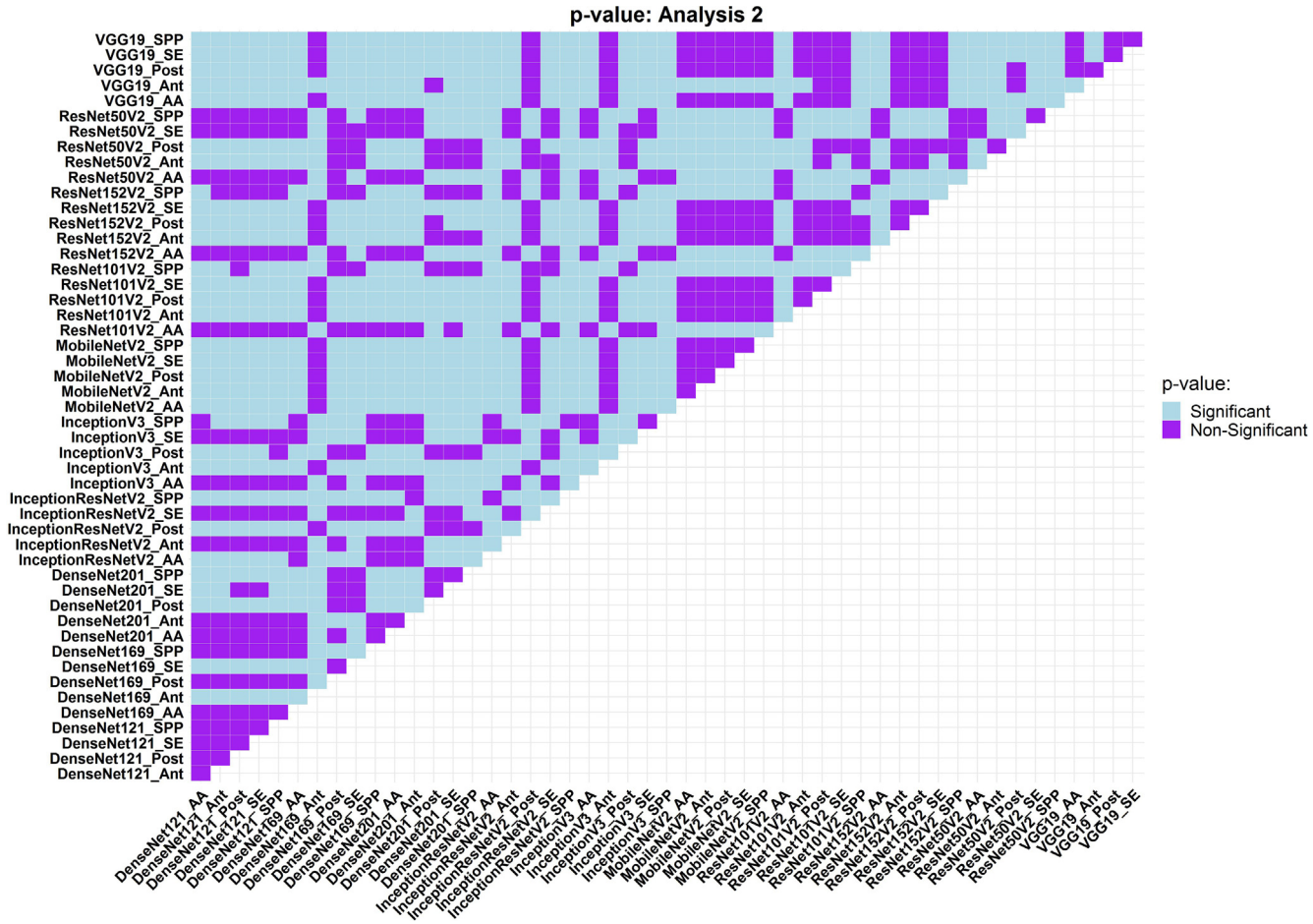

Figure 9. Model performance is compared using the DeLong test for the second strategy, which is run on the models' AUCs. The models on columns and rows were evaluated against each other. Light blue: if the row model significantly outperformed the column model in terms of p-value. Purple: if the comparison between the row model and column model yielded a non-significant p-value. Ant: anterior, Post: posterior, SPP: spatial pyramid pooling, SE: squeeze-and-excitation, AA: attention-augmented.

Table 3
Summary of DeLong test results of the top five models in the first strategy.

Models	Significantly higher than *N* models
InceptionV3_Ant	37
ResNet50V2_Post	32
DenseNet121_AA	43
InceptionResNetV2_SE	35
InceptionV3_ SPP	39

Table 4
Summary of DeLong test results of the top five models in the second strategy.

Models	Significantly higher than *N* models
DenseNet201_Ant	30
DenseNet121_Post	28
InceptionResNetV2_AA	42
DenseNet121_SE	29
InceptionResNetV2_SPP	46

tively. In addition, our best model in distinguishing normal patients from abnormal ones was DenseNet121_AA, which achieved an AUC, accuracy, sensitivity, and specificity of 72%, 72%, 73%, and 70%, respectively.

In our study, we developed two different sets of models for two different analyses strategies of WBS images while evaluating the performance of different DL algorithms, different input modes (single- and multi-view), and different aggregating methods for dual input modes, resulting in a total of 50 models for each analysis. We compare the performance of these models with human observers using different metrics. It should also be noted that the AA method for input aggregation was not evaluated in previous studies related to

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

Table 5
Average performance (AUC) of AI models for various CNN and input methods.

Method		**Analysis 1**	**Analysis 2**
Deep learning method	DenseNet121	0.559	0.643
	DenseNet169	0.575	0.604
	DenseNet201	0.556	0.611
	InceptionResNetV2	0.663	0.641
	InceptionV3	0.666	0.618
	MobileNetV2	0.500	0.500
	ResNet101V2	0.610	0.546
	ResNet152V2	0.562	0.560
	ResNet50V2	0.640	0.614
	VGG19	0.518	0.505
Input	AA	0.661	0.649
	Ant	0.578	0.580
	Post	0.582	0.573
	SE	0.620	0.604
	SPP	0.590	0.625

Table 6
Comparison of the performance of three nuclear medicine physicians (NMP) vs. artificial intelligence (AI) models* with the highest AUC in each analysis.

Analysis		TN	TP	FN	FP	ACC	AUC	SEN	SPE	Time
Analysis 1 (Normal vs. Abnormal)	NMP1	111	430	166	48	0.72	0.65	0.72	0.70	180 min
	NMP2	97	445	33	180	0.72	0.64	0.93	0.35	110 min
	NMP3	77	436	200	42	0.68	0.60	0.69	0.65	210 min
	AI	195	348	130	82	0.72	0.72	0.73	0.70	28 sec
Analysis 2 (Malignant Vs. Non-neoplastic)	NMP1	198	135	23	94	0.74	0.74	0.85	0.68	112 min
	NMP2	137	180	49	84	0.70	0.70	0.79	0.62	120 min
	NMP3	189	156	32	73	0.77	0.77	0.83	0.72	180 min
	AI	195	126	103	26	0.71	0.72	0.55	0.88	25 sec

* TN: true negative, TP: true positive, FN: false negative, FP: false positive, ACC: Accuracy, AUC: area under the ROC curve, SEN: sensitivity, SPE: specificity. The AI model in analysis 1 and 2 was DenseNet121_AA and InceptionResNetV2_SPP, respectively.

Table 7
DeLong p-values for the comparison of ROC curves between three nuclear medicine physicians (NMP) and artificial intelligence (AI) models* with the highest AUC for each analysis.

Analysis	Reader 1	Reader 2	DeLong p-value
Analysis 1 (Normal vs. Abnormal)	NMP1	AI	0.007
	NMP2	AI	0.003
	NMP3	AI	<0.001
	NMP1	NMP2	0.54
	NMP1	NMP3	0.004
	NMP2	NMP3	0.007
Analysis 2 (Malignant Vs. Non-neoplastic)	NMP1	AI	0.41
	NMP2	AI	0.65
	NMP3	AI	0.14
	NMP1	NMP2	0.14
	NMP1	NMP3	0.28
	NMP2	NMP3	0.026

* AI model in analysis 1 and 2 was DenseNet121_AA and InceptionResNetV2_SPP, respectively.

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

Table 8
Comparison of true classification between three nuclear medicine physicians (NMPs) and artificial intelligence (AI) models* for the two analysis strategies.

Analysis	Classification Type	NMP1	NMP2	NMP3	AI
Analysis 1 (Normal vs. Abnormal)	Normal	111/277	97/277	77/277	195/277
	Degenerative	117/128	117/128	119/128	95/128
	Inflammation	30/45	34/45	30/45	23/45
	Osteomyelitis	3/4	4/4	3/4	1/4
	Trauma	46/57	52/57	50/57	43/57
	Metastasis	203/210	206/210	203/210	170/210
	Tumor	31/34	32/34	31/34	16/34
Analysis 2 (Malignant vs. Non-neoplastic)	Degenerative	98/105	74/105	97/105	97/105
	Inflammation	41/44	26/44	39/44	42/44
	Osteomyelitis	9/12	5/12	9/12	11/12
	Trauma	50/60	32/60	44/60	45/60
	Metastasis	125/203	164/203	140/203	115/203
	Tumor	10/26	16/26	16/26	11/26

* AI model in analyses 1 and 2 were DenseNet121_AA and InceptionResNetV2_SPP, respectively.

the diagnosis of malignant bone diseases. In this respect, this study contributed an innovative approach toward finding the optimum model. In addition, it brings up for the first time a promising model designed to diagnose malignant bone diseases or non-neoplastic disorders. In clinical routine, NMPs use different information for bone scintigraphy assessment, including history, lab tests, pathology, and current WBS with extra spots and SPECT/CT images. In this and previous studies, only images were used as input for DL models, which do not contain all information needed for clinical diagnosis. In our study, DL models achieved comparable results with human observers' performance for two different tasks. Including the above-mentioned information is mandatory as it cannot be inferred from bone scintigraphy images alone.

In the first analysis, the DenseNet121_AA model had the highest number of correctly classified normal subjects (195/277 reflecting correctly classified as normal/total number of normal cases), while this number was lower for the three NMPs (111, 97, and 77 for NMP1-3, respectively). Other AI models resulted in lower performance compared to the three NMPs. In the second analysis, the InceptionResNetV2_SPP model in inflammation (42/44) and osteomyelitis (11/12) had superior true classification than three NMPs (41, 26, 39, and 9, 5, 9 for NMP1-3, respectively). In degenerative and trauma diseases, the AI model (97/105 and 45/60, respectively) showed higher true classifications than NMP2 (74/105 and 32/60, respectively) but was almost equal to NMP1 (98/105 and 50/60, respectively) and NMP3 (97/105 and 44/60, respectively). Unlikely, for metastases and various tumour types, the AI model had lower true classifications (115/203, 11/26) than NMP1 (125/203, 10/26), NMP2 (164/203, 16/26) and NMP3 (140/203, 16/26), respectively. DL models performed better in both analysis strategies than the three NMPs for predicting normal and non-neoplastic disorders. Nevertheless, the performance decreased for tumours and malignant bone diseases. However, AI outperformed humans' performance in the first strategy. The computational time for AI was 28 seconds, while NMPs spent 180, 110, and 210 minutes examining the same images. For the second analysis, three NMPs spent 112, 120, and 180 minutes to diagnose the 450 patients, while the designed AI model achieved comparable results in only 25 seconds. Hence, the proposed models can be utilized to reduce the workload of NMPs and save time in routine clinical nuclear medicine. In this study, there were cases that NMPs could not correctly diagnose. Yet, some of them were correctly diagnosed by AI. There were also cases where AI was not capable of detecting malignancies, some of which were correctly detected by NMPs. It is hoped that NMPs take advantage of AI tools to assist the decision-making process. AI and NMP could also complement each other to improve the overall performance.

The major limitation of our study was incomplete patient demographics and lack of additional information, such as three-phase imaging, static views, and SPECT/CT images. Furthermore, WBS scans are always evaluated in the context of correlated laboratory and clinical data; while we first exclude low quality images, the approach shown here is exclusively concerned with images, which can lead to possible artefacts and partial dosage extravasation issues. Further studies should be performed considering this additional information to improve AI models' accuracy. Large data sets could be enrolled from multiple centers using the federated learning concept, which addresses privacy issues in data sharing [48,49]. It is suggested that in future studies, physi-

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

cian's performance should be compared to outcomes of AI models to see if AI can improve physicians' performance as an assistant. In this study, only Technetium 99m-methyl diphosphonate (^{99m}Tc-MDP) was used. It is suggested to use other radiopharmaceuticals in future studies to increase the generalizability and robustness of the developed AI models. When datasets from different centers acquired with different protocols are used, it is recommended to harmonize them before applying them to the models to avoid potential errors in the DL process.

5 Conclusion

In this work, DL methods achieved promising results for the classification of WBS scans into normal and abnormal and the discrimination between malignant bone diseases and non-neoplastic disorders. In fact, by using DL algorithms, comparable results to humans are obtained without the need for expertise and experience, while the diagnostic time in DL is only a few seconds, which can significantly reduce a physician's workload. Furthermore, our models can evaluate a huge number of images without the requirement for segmentation and preprocessing on individual scans. Using the models designed in this study, a positive step can be taken toward reducing WBS interpretation time and improving the accuracy of clinical diagnosis. Furthermore, improving the accuracy of AI models through feeding them with larger and more diverse datasets should enable their clinical adoption to assist nuclear medicine physicians in the routine interpretation of WBS scans.

Disclosure

No potential conflicts of interest relevant to this article exist.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Swiss National Science Foundation under grant SNRF 320030_176052.

References

[1] Van den Wyngaert T, Strobel K, Kampen WU, Kuwert T, van der Bruggen W, Mohan HK, et al. The EANM practice guidelines for bone scintigraphy. Eur J Nucl Med Mol Imaging 2016;43:1723–1738.
[2] O'Connor MK, Brown ML, Hung JC, Hayostek RJ. The art of bone scintigraphy–technical aspects. J Nucl Med 1991;32:2332–2341.
[3] Savelli G, Maffioli L, Maccauro M, De Deckere E, Bombardieri E. Bone scintigraphy and the added value of SPECT (single photon emission tomography) in detecting skeletal lesions. Q J Nucl Med 2001;45:27–37.
[4] Fogelman I, Gnanasegaran G. Van der Wall H. Radionuclide and hybrid bone imaging. Springer; 2013.
[5] Ryan P, Fogelman I. Bone scintigraphy in metabolic bone disease. Semin Nucl Med 1997:291–305.
[6] Mari C, Catafau A, Carrio I. Bone scintigraphy and metabolic disorders. Q J Nucl Med Mol Imaging 1999;43:259.
[7] Abdelrazek S, Szumowski P, Rogowski F, Kociura-Sawicka A, Mojsak M, Szorc M. Bone scan in metabolic bone diseases. Rev Nucl Med Rev Cent East Eur 2012;15:124–131.
[8] Mundy GR. Mechanisms of bone metastasis. Cancer 1997;80:1546–1556.
[9] Peyruchaud O. Mechanisms of bone metastasis formation. J Soc Biol 2007;201:229–236.
[10] Yin JJ, Pollock CB, Kelly K. Mechanisms of cancer metastasis to the bone. Cell Res 2005;15:57–62.
[11] Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. Cancer Treat Rev 2001;27:165–176.
[12] Kimura T. Multidisciplinary Approach for Bone Metastasis: A Review. Cancers (Basel) 2018;10.
[13] Łukaszewski B, Nazar J, Goch M, Łukaszewska M, Stępi ski A, Jurczyk MU. Diagnostic methods for detection of bone metastases. Contemp Oncol (Pozn) 2017;21:98–103.
[14] Papandrianos N, Papageorgiou E, Anagnostis A, Feleki A. A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans. Appl Sci 2020;10:997.
[15] Chang CY, Gill CM, Joseph Simeone F, Taneja AK, Huang AJ, Torriani M, et al. Comparison of the diagnostic accuracy of 99 m-Tc-MDP bone scintigraphy and 18 F-FDG PET/CT for the detection of skeletal metastases. Acta Radiol 2016;57:58–65.
[16] O'Sullivan GJ, Carty FL, Cronin CG. Imaging of bone metastasis: An update. World J Radiol 2015;7:202–211.
[17] Papandrianos N, Papageorgiou E, Anagnostis A, Papageorgiou K. Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application. PLoS One 2020;15:e0237213.
[18] Pi Y, Zhao Z, Xiang Y, Li Y, Cai H, Yi Z. Automated diagnosis of bone metastasis based on multi-view bone scans using attention-augmented deep neural networks. Med Image Anal 2020;65:101784.
[19] Zhao Z, Li L, Li FL. Radiography, bone scintigraphy, SPECT/CT and MRI of fibrous dysplasia of the third lumbar vertebra. Clin Nucl Med 2009;34:898–901.
[20] Bombardieri E, Aktolun C, Baum RP, Bishof-Delaloye A, Buscombe J, Chatal JF, et al. Bone scintigraphy: procedure guidelines for tumour imaging. Eur J Nucl Med Mol Imaging 2003;30:Bp99-106..
[21] Sanaat A, Akhavanalaf A, Shiri I, Salimi Y, Arabi H, Zaidi H. Deep-TOF-PET: Deep learning-guided generation of time-of-flight from non-TOF brain PET images in the image and projection domains. Hum Brain Mapp 2022;43:5032–5043.
[22] Sanaat A, Shooli H, Ferdowsi S, Shiri I, Arabi H, Zaidi H. DeepTOFSino: A deep learning model for synthesizing full-dose time-of-flight bin sinograms from their corresponding low-dose sinograms. Neuroimage 2021;245:118697.
[23] Shiri I, Arabi H, Sanaat A, Jenabi E, Becker M, Zaidi H. Fully Automated Gross Tumor Volume Delineation From PET in Head and Neck Cancer Using Deep Learning Algorithms. Clin Nucl Med 2021;46:872–883.

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008

[24] Salimi Y, Shiri I, Akhavanallaf A, Mansouri Z, Saberi Manesh A, Sanaat A, et al. Deep learning-based fully automated Z-axis coverage range definition from scout scans to eliminate overscanning in chest CT imaging. Insights Imaging 2021;12:162.
[25] Akhavanallaf A, Mohammadi R, Shiri I, Salimi Y, Arabi H, Zaidi H. Personalized brachytherapy dose reconstruction using deep learning. Comput Biol Med 2021;136:104755.
[26] Jabbarpour A, Mahdavi SR, Vafaei Sadr A, Esmaili G, Shiri I, Zaidi H. Unsupervised pseudo CT generation using heterogenous multicentric CT/MR images and CycleGAN: Dosimetric assessment for 3D conformal radiotherapy. Comput Biol Med 2022;143:105277.
[27] Papandrianos NI, Papageorgiou EI, Anagnostis A, Papageorgiou K, Feleki A, Bochtis D. Development of Convolutional Neural Networkbased models for bone metastasis classification in nuclear medicine. In: 11th International Conference on Information, Intelligence, Systems and Applications (IISA: IEEE. p. 1–8.
[28] Zhao Z, Pi Y, Jiang L, Xiang Y, Wei J, Yang P, et al. Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. Sci Rep 2020;10:17046.
[29] Khodabakhshi Z, Shiri I, Zaidi H, Andratschke N, Tanadini-Lang S . Two-Year Overall Survival Prediction in Non–Small-Cell Lung Cancer Patients Using Pre-Treatment Computed Tomography Images and Deep Neural Networks: A Multicentric Study. Proceedings of Medical Imaging with Deep Learning (MIDL 2022), 2022.
[30] Aoki Y, Nakayama M, Nomura K, Tomita Y, Nakajima K, Yamashina M, et al. The utility of a deep learning-based algorithm for bone scintigraphy in patient with prostate cancer. Ann Nucl Med 2020;34:926–931.
[31] Han S, Oh JS, Lee JJ. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. Eur J Nucl Med Mol Imaging 2022;49:585–595.
[32] Hsieh T-C, Liao C-W, Lai Y-C, Law K-M, Chan P-K, Kao C-H. Detection of Bone Metastases on Bone Scans through Image Classification with Contrastive Learning. J Pers Med 2021;11:1248.
[33] Liu S, Feng M, Qiao T, Cai H, Xu K, Yu X, et al. Deep Learning for the Automatic Diagnosis and Analysis of Bone Metastasis on Bone Scintigrams. Cancer Manag Res 2022;14:51.
[34] Jouirou A, Baâzaoui A, Barhoumi W. Multi-view information fusion in mammograms: A comprehensive overview. Information Fusion 2019;52:308–321.
[35] Carneiro G, Nascimento J, Bradley AP. Automated Analysis of Unregistered Multi-View Mammograms With Deep Learning. IEEE Trans Med Imaging 2017;36:2355–2365.
[36] Khan HN, Shahid AR, Raza B, Dar AH, Alquhayz H. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access 2019;7:165724–165733.
[37] Shuo W, Mu Z, Gevaert O, Zhenchao T, Di D, Zhenyu L, et al. A multi-view deep convolutional neural networks for lung nodule segmentation. Annu Int Conf IEEE Eng Med Biol Soc 2017;2017:1752–1755.
[38] Liu X, Hou F, Qin H, Hao A. Multi-view multi-scale CNNs for lung nodule type classification from CT images. Pattern Recognit 2018;77:262–275.
[39] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2020;42:2011–2023.
[40] He K, Zhang X, Ren S, Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell 2015;37:1904–1916.
[41] Bello I, Zoph B, Vaswani A, Shlens J, Le QV. Attention augmented convolutional networks. Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 3286–95.
[42] Erdi YE, Humm JL, Imbriaco M, Yeung H, Larson SM. Quantitative bone metastases analysis based on image segmentation. J Nucl Med 1997;38:1401–1406.
[43] Yin TK, Chiu NT. A computer-aided diagnosis for locating abnormalities in bone scintigraphy by a fuzzy system with a three-step minimization approach. IEEE Trans Med Imaging 2004;23:639–654.
[44] Huang J-Y, Kao P-F, Chen Y-S. A set of image processing algorithms for computer-aided diagnosis in nuclear medicine whole body bone scan images. IEEE Trans Nucl Sci 2007;54:514–522.
[45] Geng S, Ma J, Niu X, Jia S, Qiao Y, Yang J. A mil-based interactive approach for hotspot segmentation from bone scintigraphy. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): IEEE; 2016. p. 942–6.
[46] Sadik M, Jakobsson D, Olofsson F, Ohlsson M, Suurkula M, Edenbrandt L. A new computer-based decision-support system for the interpretation of bone scans. Nucl Med Commun 2006;27:417–423.
[47] Sadik M, Hamadeh I, Nordblom P, Suurkula M, Höglund P, Ohlsson M, et al. Computer-assisted interpretation of planar whole-body bone scans. J Nucl Med 2008;49:1958–1965.
[48] Shiri I, Vafaei Sadr A, Akhavan A, Salimi Y, Sanaat A, Amini M, et al. Decentralized collaborative multi-institutional PET attenuation and scatter correction using federated deep learning. Eur J Nucl Med Mol Imaging 2022;1–17.
[49] Shiri I, Vafaei Sadr A, Amini M, Salimi Y, Sanaat A, Akhavanallaf A, et al. Decentralized Distributed Multi-institutional PET Image Segmentation Using a Federated Deep Learning Framework. Clin Nucl Med 2022;47:606–617.

Please cite this article as: G. Hajianfar, M. Sabouri, Y. Salimi et al., Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal deep learning algorithm and comparison with human observer performance, Z Med Phys, https://doi.org/10.1016/j.zemedi.2023.01.008