Deep Learning Classifiers to Reduce False Positives in Osteolytic Lesion Segmentation Results from Low-dose CT Scans of Multiple Myeloma

Munirdin Jadikar[1*], Martijn van Leeuwen[1*], Thijs van Oudheusden[2], Sebastian Oei[2], Bastiaan Steunenberg[2], Rik Kint[2], Erik R. Ranschaert[3], Gerlof P.T. Bosma[2], Gorkem Saygili[1], and L.L. Sharon Ong[1]

[1] Dept. of CSAI, Tilburg University, Tilburg, the Netherlands
{m.p.vanleeuwen,m.jadikar,g.saygili,l.l.ong}@tilburguniversity.edu
[2] Elisabeth-TweeSteden Ziekenhuis, Tilburg, the Netherlands.
{t.vanoudheusden,b.oei,b.steunenberg,r.kint,gpt.bosma}@etz.nl
[3] St. Nikolaus Hospital,Eupen and Ghent University, Belgium
{erik.ranschaert}@ugent.be

Abstract. Multiple myeloma (MM) is a hematological malignancy with a low survival rate if not detected in an early stage. A common symptom of MM is the development of osteolytic lesions, which appear as hypodense regions in bone tissue that are often visualised using low-dose CT imaging. Finding one lesion with a diamater of 5 mm or more is already enough to support the diagnosis of MM. However, evaluation of total-body CT (TBCT) scans is time consuming. Our group has developed an automated lesion segmentation algorithm to assist this process. Although providing accurate detection results, the algorithm results in excessive lesion-like false positive candidates. To address this problem and further improve the segmentation performance, we deployed deep learning classifiers to reduce the false positive rate as a post-processing step. To train and evaluate the classifiers, a dataset was created, comprising of patches of lesions annotated by radiologists and images patches containing healthy bone tissue. The results showed that the best performing model, a fine-tuned ResNet50 model, achieved an F1 score of 0.83 on the test set. To test the performance of the model, expert radiologists labelled segmentation results as true or false positives for a hold-out test set. The model achieved an F1 score of 0.68 and a False Positive Rate (FPR) of 0.47 on the hold-out test set, reducing the number of false positives by 53%. By integrating our proposed model to the original segmentation pipeline, the number of reported false positives can be reduced, leading to a more reliable system.

Keywords: False positive reduction · Osteolytic lesion segmentation · Convolutional Neural Networks

[*] These authors contributed equally to this work

1 Introduction

Multiple myeloma (MM) is a hematological malignancy which leads to an uncontrolled growth of plasma cells [2]. MM patients often display symptoms such as hypercalcemia, anemia and the development of osteolytic lesions [2]. Between 80% to 90% of MM patients develop these lesions during the course of their disease [9]. One of the imaging methods that is commonly used to examine osteolytic lesions is low-dose Computed Tomography (CT) imaging of the whole body [17].

CT scans provide radiologists with valuable information to visually detect and measure bone lesions, which appear as small holes in bone tissue [10]. Furthermore, osteolytic lesions are small and can be found in multiple locations in the body. However, in cross-sectional imaging, the interpretation errors lie between 20%-30% [6]. In addition, radiologists must possess the expertise to detect osteolytic lesions, making it more difficult for inexperienced radiologists to detect them [4]. Because of these reasons, there is a possibility that radiologists overlook certain lesions when evaluating a patient's CT scan. According to the International Myeloma Working Group (IMWG), a patient can be diagnosed with MM if at least one lesion with a diameter of $\geq 5\ mm$ is present [7]. Therefore, detecting lesions is crucial for the diagnosis and treatment of the patient.

There already exist a number of deep learning models that focus on the automated detection of osteolytic lesions [17][1]. We have also developed a deep learning model to segment osteolytic lesions in low-dose CT-scans. However, due to the similarity of osteolytic lesions with other hypodense regions in bone tissue, our segmentation model was prone to produce false positive predictions. Figure 1 shows a number of examples of false positive predictions found by this segmentation model. The highlighted regions show the segmentation results which were confirmed to be false positive predictions by radiologists.

To improve the reliability and adoption of a lesion detection/segmentation algorithm without compromising the segmentation accuracy, we propose a post-processing strategy to reduce the false positive predictions. By classifying the output of the segmentation results as lesions or non-lesions, it is possible to determine whether a potential lesion site is an actual lesion or a false positive prediction. This post-processing model was developed by fine tuning pre-trained classification models on 96 low-dose CT-scans acquired at Elisabeth-TweeSteden hospital (ETZ) in Tilburg. This study is the first to investigate a classification model in the post-processing of the bone lesion segmentation model to classify false positive predictions. Additionally, the techniques employed in this study, such as TL and ensemble learning with limited data, can be applied to other medical classification tasks as well. Ultimately, the findings of this study will display the value of using an additional post-processing model to reduce false positive predictions produced by a segmentation/detection model.

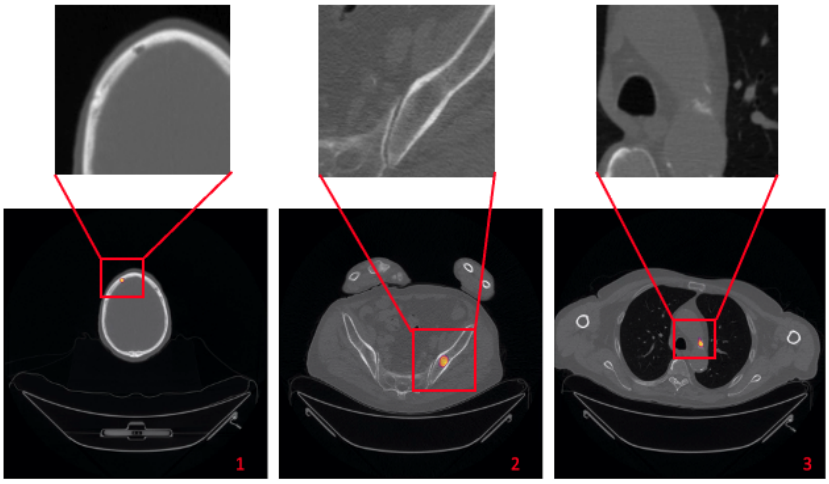

Fig. 1. False positives in segmented bone lesions. 1: Granulation, 2: Yellow Bone Marrow, 3: Calcified Aorta.

2 Related Work on False Positive Detection in Medical Image Analysis

There is limited work on classification of false positives in medical image analysis using deep learning, with no known work for classifying osteolytic lesion segmentation results. These methods normally apply a two stage approach where the first stage is an object detection or segmentation algorithm, and the second stage is a classifier, typically DCNN-based.

Multiple applications of this strategy have been developed on to the LIDC-IDRI dataset aiming to reduce false positives in pulmonary lung nodule detection. In [13], a multi-view CNN was proposed to classify pulmonary nodules. A set of 2D patches at different cross sections and different orientations were extracted for each class. Their proposed architecture consisted of multiple streams of 2D ConvNets, of which the outputs were combined using a reliable fusion method to get the final classification. In [20], multi-scale 2D CNNs were proposed to reduce the number of false positive predictions, which was also applied on automated pulmonary nodule detection. That proposed method used three different 2D images cropped from 3D CT scans to preserve spatial information and shorten training time, as 2D CNN are more computationally efficient compared to 3D.

Another alternative to reduce false positive rates, is applying an ensemble classifier which was proposed in [19]. This particular ensemble classifier was used to classify false positive brain metastases segmentation results. This classifier was composed of a Siamese network and a support vector machine (SVM) classifier, and was designed to reduce the false positive rate of the segmentation results.

There is no known previous work done on using a classification model to reduce false positive predictions made by a osteolytic lesion segmentation model. The primary objective of this study is to investigate the application of pre-trained CNN models to classify false positives from a osteolytic lesion segmentation model.

3 Experimental Setup

Figure 2 shows the experimental setup used in this paper. Four different pre-trained models; VGG16, InceptionV3, ResNet50 and EfficientNetB7 were explored. A description of these models and the preprocessing applied to each model is presented in Appendix 1.The first step in training these models is the preprocessing of the data where image patches with and without lesions were extracted. Due to the limited number of labeled data, data augmentation was applied to increase the training dataset size. Using this augmented dataset, the pre-trained models were fine-tuned by unfreezing all the hidden layers. After finishing training the models, all models were evaluated on a test set and a separate hold-out test set.

3.1 Dataset and Data Preparation

The dataset consists of 96 CT scans from 79 patients diagnosed with multiple myeloma from Elisabeth-TweeSteden Hospital (ETZ) in Tilburg. The dataset is

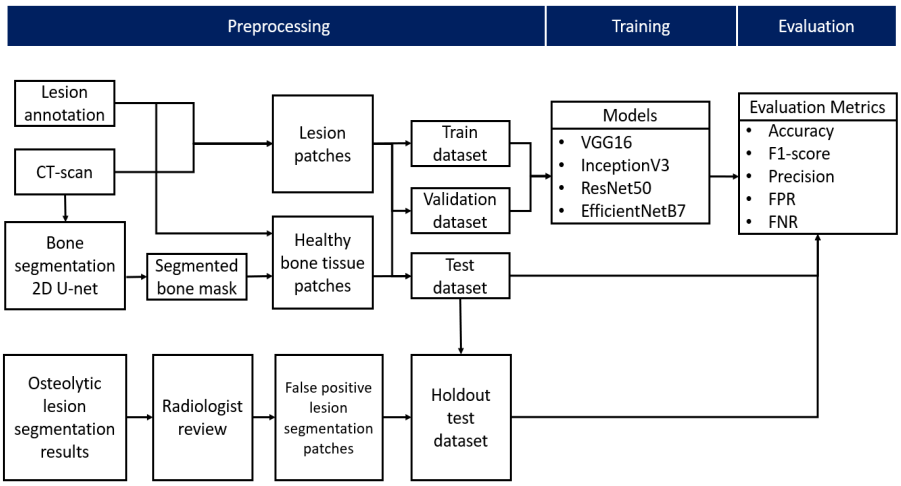

Fig. 2. Experimental workflow. First, images patches are created for training, validation and testing. Four pre-trained models are tested and evaluated. Finally, the models are tested on a hold out test set. The hold out test set was created from segmentation results which were reviewed by radiologists.

anonymized for all demographic information. Each patient had a full body scan or a combination of an upper and lower body scan. The average number of lesions per patient was four, with an interquatile range of 1.5 to 8.5 lesions. Each axial slice had a resolution of either 512×512 pixels or 768×768 pixels, and the slice thickness was either 2.5 mm or 3 mm. The lesions were annotated by radiologists from the ETZ hospital by creating segmentation masks with 3D slicer. In the 96 low-dose CT-scans, 665 lesions were annotated. From these lesions, 2D patches of 224×224 pixels, were made by randomly cropping around the centroid of the lesions. This approach allowed the lesions to be located in random positions in the patches. The extracted patches were split (subject-wise) into a training, validation, and test set.

The classifiers should learn to distinguish lesion from non-lesion tissue. Since these lesions are predominantly present in bone tissue, patches of healthy bone tissue representing a negative class were deemed suitable. However, extracting a lesion-free patch at a random location in the CT scan can result in a patch in which no bone tissue is present. To ensure that bone tissue is present in these patches, a bone tissue mask was made for each scan using a 2D U-Net segmentation model [12], trained on the CT-ORG dataset [11]. The bone masks were then used to randomly generate patches, in which at least 10% of the patch consisted of bone tissue, while ensuring that no annotated lesions were present in the patch.

To create a hold-out test set, a U-Net segmentation algorithm was trained on the lesion annotation masks that were made by ETZ radiologists. This model was trained on patches of 192x192 pixels and tested on a number of low-dose CT-scans using a sliding window approach [8]. These results were then evaluated by a group of radiologists who labelled the results as true positives or false positives. This labeled data will be referred to as the hold-out test set throughout this paper. For this hold-out test set, the positive lesion patches were randomly sampled from the positive patches in the original test set. Table 1 shows an overview of the number of patches/patients in each dataset.

Table 1. The number of image patches and patients in each dataset

	No. Lesion patches	No. Non-lesion patches	No. Patients
Training data	5267	5267	59
Validation data	710	710	10
Test data	705	705	10
Hold-out test data	227	227	4 (from test set)

3.2 Model Design

The standard procedure for applying transfer learning involves the removal of the final classification layers of a pre-trained model and replacing them with a self-designed classifier. This pre-trained model, called the base model, is utilized as a feature extractor. Commonly, the base model is followed by a Global Average Pooling (GAP) layer or a flatten layer. However, the use of a flatten layer increases susceptibility to overfitting, particularly when the amount of training data is limited [5]. The GAP layer on the other hand mitigates the overfitting by minimizing the overall number of parameters in the model. As there is a limited amount of data, the GAP layer was chosen over a flatten layer to limit the chance of overfitting. After removal of the final classification layer and addition of the GAP layer, a dense layer and a final output layer were added to the pre-trained models as displayed in Figure 3. The final output layer is a sigmoid activation

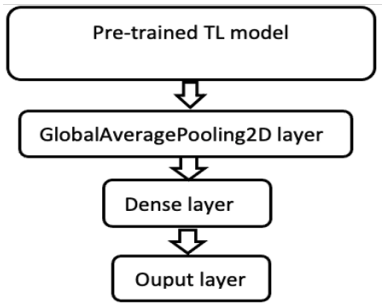

Fig. 3. Model design with base models

function. For more information on the hyperparameters tuned, see Appendix C.

3.3 Evaluation Metrics

To evaluate our models, we computed the accuracy, precision and F1 score which are typically used for classification tasks. As our focus is on the removal of false positives, we also compute the False Positive Ratio (FPR) and False Negative Rate(FNR) for model evaluation. These metrics can calculated from sensitivity/recall and specificity. Here, $FPR = 1 - specificity$, and $FNR = 1 - Recall$. To visualize the trade-off between true and false positives, we computed the ROC curve.

4 Results

In the following section shows the evaluation of the models on both test sets and shows some examples of correct and incorrect classifications.

4.1 Evaluation of Models on Test Set

Table 2 shows the test accuracy for each model. The results indicated that fine-tuned ResNet50 model outperformed the other models, reaching a test accuracy of 0.84, a precision score of 0.86, and an F1-score of 0.83. The ResNet50 model achieved an FPR of 0.12 and an FNR of 0.20. Figure 4 shows the confusion matrix and ROC curve of the ResNet50. The number of FPs and FNs are relatively low, and the ROC curve reveals that both classes have a high AOC (0.91)

Table 2. The results of fine-tuned models on the test set

Models	Test Accuracy	Precision	F1-score	FPR	FNR
VGG16	0.73	0.74	0.73	0.26	0.28
InceptionV3	0.81	0.80	0.81	0.21	0.18
ResNet50	**0.84**	**0.86**	**0.83**	**0.12**	**0.20**
EfficientNetB7	0.77	0.83	0.75	0.14	0.31

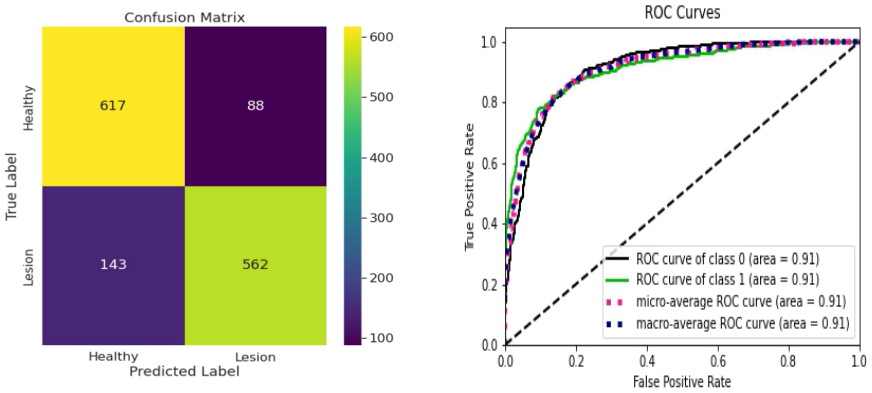

Fig. 4. Confusion Matrix and ROC curve of ResNet50 on the test dataset

4.2 Generalization with Hold-out Test Set

Table 3 displays the final result of the four models on the hold-out test set. Overall, the ResNet50 and InceptionV3 performed better than VGG16 and EfficientNetB7. The ResNet50 model was the best performing model with a test

accuracy of 0.64, a precision of 0.62 and a F1-score of 0.68. However, classification performance decreased on the hold-out test set compared to the previous test set.

Table 3. The results of fine-tuned models on the hold-out test set

Models	Test Accuracy	Precision	F1-score	FPR	FNR
VGG16	0.52	0.51	0.59	0.65	0.31
InceptionV3	0.64	0.60	0.70	0.55	0.16
ResNet50	**0.64**	**0.62**	**0.68**	**0.47**	**0.24**
EfficientNetB7	0.58	0.57	0.61	0.51	0.33

Figure 5 shows the confusion matrix for the ResNet50. There were 107 of 227 non-lesions classified as lesions. The ROC curves in Figure 5 also show an AUC of 0.70 which is significantly lower than the earlier observed 0.90 in Figure 4. Overall, the model on the hold-out test set did not perform as well as it did on the test set that contained automatically generated healthy bone tissue patches. This may be attributed to the method employed for extracting the patches. The healthy bone tissue patches from the test dataset were extracted based on the percentage of bone in the patch, whereas the false positive feedback samples in the hold out test set, were already more difficult examples of bone tissue that shows resemblance with osteolytic lesions. This might have led to a more difficult set of healthy bone tissue patches in the held-out test set.

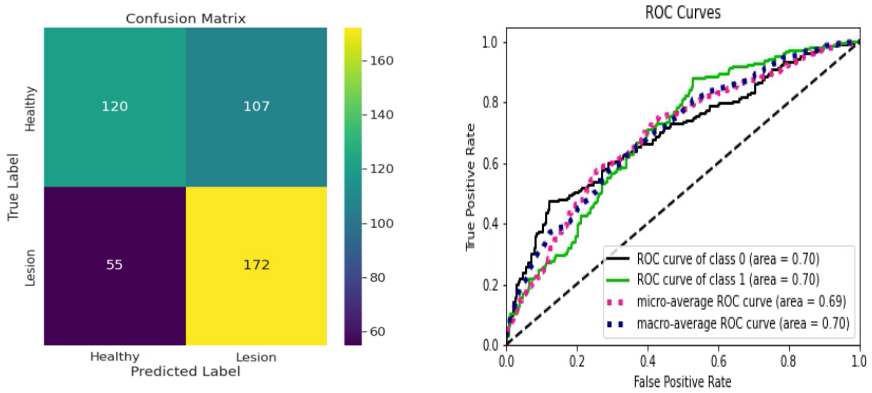

Fig. 5. Confusion matrix and ROC curve of ResNet50 on the hold-out test set

Figure 6 and Figure 7 show a number of examples of correct and incorrect classifications. The top two images in Figure 6 contain osteolytic lesions which were misclassified as healthy bone tissue and the bottom two images display

healthy bone tissues that were misclassified as bone lesions. Figure 7 shows a number of examples of correct classifications made by ResNet50. It shows a correct classification of a lesion in the skull and vertebra, and a correct classification of healthy bone tissue in the pelvis and skull.

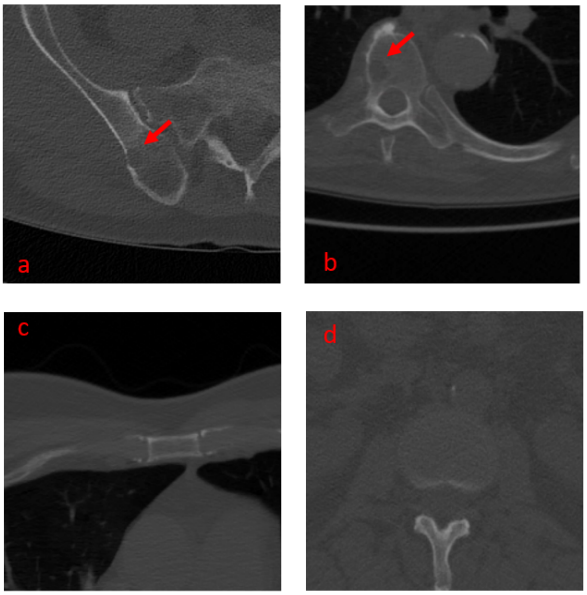

Fig. 6. Example images which were misclassified by ResNet50 as false negatives (top) and false positives (bottom). Patch *a* and patch *b* both contain a lesion but were incorrectly labeled as healthy bone tissue by ResNet50. Patch *a* shows a lesion in the pelvis patch *b* displays a lesion in the vertebra. Patch *c* displays costal cartilage between the sternum and the ribs and patch *d* shows an intervertebral disc. Both these regions have similarities with osteolytic lesions which is presumable the reason why these are classified as lesions by ResNet50.

5 Discussion

There is a scarcity of literature about reducing false positives in bone lesion classification. At the same time, no studies have been done on classifying false positives from segmented bone lesions. Additionally, only a few studies have focused on enhancing the quantity and quality of medical datasets. While TL on medical images has been extensively studied, there is limited research on applying it on the classification of segmented bone lesions. This paper displayed an application of TL of classification models on a clinically relevant problem.

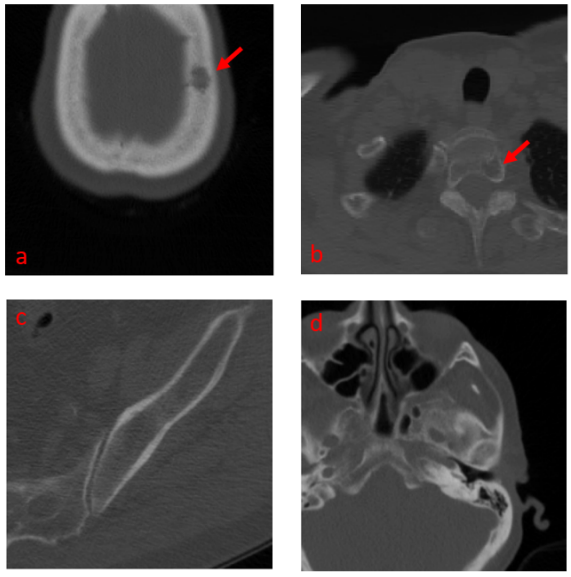

Fig. 7. Example images which were correctly classified by ResNet50 as true positives (top) and true negatives (bottom). Patch *a* and *b* show a lesion the skull and a vertebra respectively, which were both correctly classified by ResNet50. Patch *c* shows a hypodense region in the pelvis which was correctly identified as healthy tissue, and also the complex bone tissue in the skull shown in patch *d* was correctly labeled as healthy bone tissue.

The proposed dataset performed well in distinguishing automatically generated bone tissue patches from patches that contained a osteolytic lesion. However, the model performance seemed to drop when it was applied to a dataset composed of false positive osteolytic lesion segmentation results. The proposed ResNet50 classification model was able to detect 120 out of the 227 false positive segmentations. The reduction of false positives by only 53% could be partially attributed to the method in which the training, validation and test dataset was generated. We expect that the model performance will increase when the model is not only evaluated on a dataset of false positive segmentation results, but also trained on this data. Furthermore, the automated generation of healthy bone tissue patches can lead to the inclusion of unannotated lesions in the training data which can prevent the model to find the optimal weight configuration.

A significant limitation of this study is the limited availability of annotated data. The identified lesions from CT scans are scarce and so is the availability of a relevant negative class. However, we show that we can remove more than half of the false positive segmentation results with our current set-up. To further improve this work, future work includes the expansion of our dataset, and further

exploration of the usage of pretrained models and radiomic features on reducing false positive predictions.

6 Conclusion

This is the first work to show that false positives can be reduce by a helper classification model on osteolytic lesions. To train and evaluate the classifiers, a dataset was created, comprising of image patches of lesions annotated by radiologists and images patches containing healthy bone tissue. The results showed that the best performing model, a fine-tuned ResNet50 model, achieved an F1 score of 0.83 on the test set. A group of radiologists labelled segmentation results as true or false positives for a hold-out test set on which the model achieved an F1 score of 0.68 and a False Positive Rate (FPR) of 0.47. By integrating our proposed model to the original segmentation platform, the number of false positives can be reduced, leading to a more reliable system and a reduced workload for radiologists. These outcomes suggest that it is feasible to train osteolytic lesion classifiers using pre-trained DCNN models on limited datasets. However, the final results indicate that the model is not yet robust enough and requires more research.

References

1. Faghani, S., Baffour, F.I., Ringler, M.D., Hamilton-Cave, M., Rouzrokh, P., Moassefi, M., Khosravi, B., Erickson, B.J.: A deep learning algorithm for detecting lytic bone lesions of multiple myeloma on ct. Skeletal Radiology **52**(1), 91–98 (2023)
2. Filho, A.G., Carneiro, B.C., Pastore, D., Silva, I.P., Yamashita, S.R., Consolo, F.D., Hungria, V.T., Sandes, A.F., Rizzatti, E.G., Nico, M.A.: Whole-body imaging of multiple myeloma: Diagnostic criteria. Radiographics **39**, 1077–1097 (2019)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
4. He, Y., Pan, I., Bao, B., Halsey, K., Chang, M., Liu, H., Peng, S., Sebro, R.A., Guan, J., Yi, T., Delworth, A.T., Eweje, F., States, L.J., Zhang, P.J., Zhang, Z., Wu, J., Peng, X., Bai, H.X.: Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. EBioMedicine **62**, 103121 (2020)
5. Lin, M., Chen, Q., Yan, S.: Network in network. In: International Conference on Learning Representations (ICLR) (2014)
6. Maskell, G., Frcp, F.: Commentary error in radiology-where are we now? (2019)
7. Moulopoulos, L.A., Koutoulidis, V., Hillengass, J., Zamagni, E., Aquerreta, J.D., Roche, C.L., Lentzsch, S., Moreau, P., Cavo, M., Miguel, J.S., Dimopoulos, M.A., Rajkumar, S.V., Durie, B.G.M., Terpos, E., Delorme, S.: Recommendations for acquisition, interpretation and reporting of whole body low dose CT in patients with multiple myeloma and other plasma cell disorders: a report of the IMWG Bone Working Group. Blood cancer journal **8**, 95 (2018)
8. Ong, L.S., Martijn van Leeuwen, G.S.: A deep learning-based approach to detect and segment osteolytic bone lesions in whole-body, low-dose ct imaging of multiple myeloma patients. EuSoMII Annual Meeting 2022 (2022)

9. Rajkumar, S.V., Kumar, S.: Multiple myeloma: Diagnosis and treatment. Mayo Clinic Proceedings **91**, 101–119 (1 2016)
10. Reagan, M.R., Liaw, L., Rosen, C.J., Ghobrial, I.M.: Dynamic interplay between bone and multiple myeloma: Emerging roles of the osteoblast. Bone **75**, 161–169 (6 2015)
11. Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L.: Ct-org, a new dataset for multiple organ segmentation in computed tomography. Scientific Data **7** (2020)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
13. Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sánchez, C.I., Van Ginneken, B.: Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging **35**(5), 1160–1169 (2016)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojnaw, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
16. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). vol. 97, pp. 6105–6114 (2019)
17. Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P., Piraud, M., Buck, A., Shi, K., Menze, B.H.: Automated whole-body bone lesion detection for multiple myeloma on 68 ga-pentixafor pet/ct imaging using deep learning methods. Contrast Media and Molecular Imaging **2018** (2018)
18. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging **5**, 1 (2018)
19. Yang, Z., Chen, M., Kazemimoghadam, M., Ma, L., Stojadinovic, S., Timmerman, R., Dan, T., Wardak, Z., Lu, W., Gu, X.: Deep-learning and radiomics ensemble classifier for false positive reduction in brain metastases segmentation. Physics in Medicine & Biology **67**(2), 025004 (2022)
20. Zhao, D., Liu, Y., Yin, H., Wang, Z.: A novel multi-scale cnns for false positive reduction in pulmonary nodule detection. Expert Systems with Applications p. 117652 (2022)

A Models for Transfer Learning

With limited data, we applied transfer learning for classification. Four different, well-known pretrained models were explored; VGG16, InceptionV3, ResNet50 and EfficientNetB7, described below.

A.1 VGG16

The VGG16 model consists of 13 convolutional layers with 3×3 filters (Figure 8). The convolutional layers can be divided into five blocks, and a max pooling layer follows each. The final max pooling layer is connected to a flatten layer followed

by three dense layers. The final dense layers have 1000 units with a sigmoid activation function. TL tasks without a modified VGG16 model require an input image shape of 224×224 [14]. When its final dense layers are removed, it can be modified with different dense layers with different input image shapes. Compared to many other models, the VGG16 has a shallow structure and requires less computational load [18]. Hence, it is chosen as the baseline model.

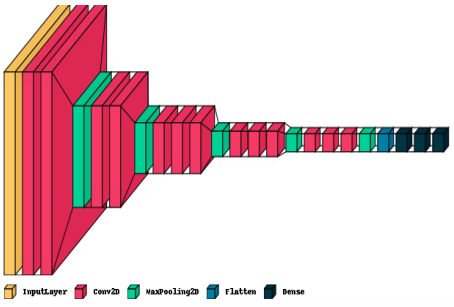

Fig. 8. VGG16 model structure. Generated from keras-visualizer

A.2 InceptionV3

The inceptionV3 architecture contains multiple inception modules (Figure 9), stacked upon each other [15]. We chose to explore this model for classification as it is robust against overfitting with limited data. Compared to previous inception models, the filters are smaller (3×3), requiring less computational power. Computation is also reduced with asymmetric convolution filters[15].

A.3 ResNet50

The Resnet50 model consisted of 50 layers [3] with residual network architectures. Very Deep learning structures suffer from gradient vanishing or exploding problems. Residual networks overcome this problem by mapping the activation to two or three layers ahead when it is added to the layer. In Resnet50, the residual block is designed in the bottleneck approach, which allows the model to train faster. The bottleneck building block contains three convolutional layers with 1×1, 3×3, and 1×1 filters, and the 1×1 filter reduces the trainable parameters [3]. Despite the deeper structure of Resnet50, it has fewer floating point operations (FLOPs) than shallow models like VGG19.

A.4 EfficientNetB7

EfficienNetB7 is one of the new versions of the EfficientNet model family. This model is designed based on a new way to scale model dimensions like depth,

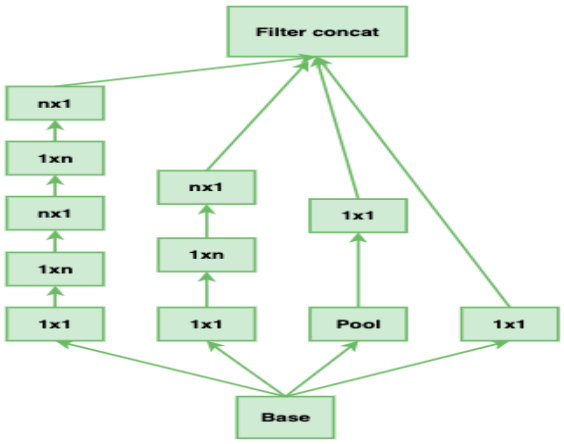

Fig. 9. Inception module with n × n convolution. Adapted from: [15]

width, and resolution [16]. First, the relationship of different scaling dimensions is searched based on the grid search to find an optimal value for the compound scaling method. Then, the model is scaled up according to the compound coefficient, balancing the network dimensions optimally (Figure 10). This model is relatively new, and there is limited study on EfficientNetB7 in medical image classification. However, it is a relatively large model with more than 66 million parameters and is expected to achieve very high accuracy on image classification [4].

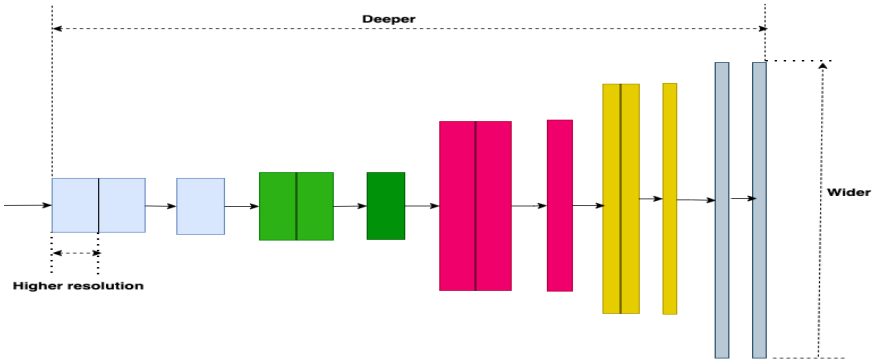

Fig. 10. Model scaling methods in EfficientNetB7. Adapted from: [16]

A.5 Requirements for Preprocessing Images in Pre-trained Models

The pre-trained models employed in this study were trained on different image input shapes. Omitting the fully-connected layers at the top of the model enables the use of different input shapes for the osteolytic lesion classification. Furthermore, each pre-trained model necessitated a specific method for processing the input images 4. These preprocessing methods allow the pre-trained models to achieve the best performance. The TensorFlow framework offers a built-in preprocessing approach for each pre-trained model. These models were trained on ImageNet and subsequently fine-tuned to enhance their performances.

Table 4. Preprocessing requirements of pre-trained models

Model name	Pre-trained image shape	Preprocessing method
VGG16	224x224	Converted to BGR and zero-centered.
InceptionV3	299x299	[-1,1]
Resnet50	224x224	Converted to BGR and zero-centered.
EfficienNetB7	224x224	[0,255]

B Data Augmentation

Table 5 shows the data augmentation techniques which were applied in this paper.

Table 5. Basic data augmentation techniques

Data augmentation type	Range
Rotation	[0,180]
Width shift	[0,1]
Height shift	[0,1]
Horizontal flip	True, False
Shear	[0,1]
Zoom	[0,1]
Fill mode	Nearest, constant, reflect, wrap

C Hyperparameter Tuning

A total of eight different hyperparameters with varying values were considered for each pre-trained model (Table 6). The models were trained with pre-trained weights and fine-tuned by unfreezing the trainable layers.

Table 6. Hyperparameters of models with selected values

Hyper-params	**Values range**	**Best option**	**Help function**
Learning rate	[0.01, 0.0000001]	Determined by ReduceLROnPlat-eau	ReduceLROnPlateau
Dense layer	Flatten () GlobalAveragePooling2D()	GlobalAveragePooling2D()	None
Units in dense layers	[32,64,128,512]	64	None
Batch size	[4,8,16,32,64]	32	None
Optimizers	Adam, Nadam	Adam	Keras. Optimizers.Adam
Loss functions	Binary cross entropy Categorical crossentropy	Binary cross entropy	None
Epochs	[10,20,40]	40	EarlyStopping
Freeze and unfreeze layers	True, False	True	model.trainable