Detection of Osteosarcoma on Bone Radiographs Using Convolutional Neural Networks

Larissa Y. Asito*, Hélcio M. Pereira†, Marcello H. Nogueira-Barbosa†, Renato Tinós*
*Department of Computing and Mathematics
University of São Paulo
Ribeirão Preto, Brazil
Email: larissa.asito@alumni.usp.br, rtinos@ffclrp.usp.br
†Department of Medical Imaging, Hematology and Clinical Oncology
University of São Paulo
Ribeirão Preto, Brazil
Email: helcioradio40@usp.br, marcello@fmrp.usp.br

***Abstract*—Osteosarcoma is the most common type of bone cancer. We propose a computer-aided diagnosis system based on convolutional neural networks (CNNs) for the identification of osteosarcoma on bone radiographs. The CNN should indicate regions of the image that may contain tumors. In order to indicate these regions on the image, we propose to split the image in windows and individually classify them by using a CNN. Techniques for pre-processing, such as window exclusion and labeling, are proposed. Two CNNs are compared in the proposed system. The first one is trained from scratch, while the second one is a pre-trained CNN (VGG16). The CNNs are compared to four machine learning models that use features extracted from the image windows as inputs: multilayer perceptron (MLP), decision tree, random forest, and MLP with feature selection. In the experiments, the best performance was obtained by the pre-trained CNN.**

***Index Terms*—Artificial neural networks, Radiography, Decision support systems**

I. INTRODUCTION

Bone tumors are usually discovered incidentally on imaging exams taken to investigate other medical problems. Many of these tumors are benign. Malignant bone tumors usually originate from metastasis of other cancers. Osteosarcoma is the most common type of primary malignant bone tumor, occurring most often in adolescents and young adults. Medical images are important for the early detection of malignant tumors. In addition, images are helpful for estimating malignancy [1] and, as a consequence, for planning treatments.

In a near future, computer-aided diagnosis systems may be used to inform the health professional about possible occurrence of osteosarcoma on routinely generated images. As there is an abundance of bone radiographs (X-ray images), such systems will be extremely useful. A characteristic often seen in these images in patients with osteosarcoma is the Codman triangle, which is basically a lesion formed when the periosteum is elevated due to the tumor [2], [3]. This characteristic, as well as others that characterize osteosarcoma, should assist in the automatic classification of the tumor.

In a scenario without automation, the radiologist must carefully examine all images to identify regions that may indicate the presence of tumors, even if the objective is to identify other occurrences in the image, e.g., fractures. Considering that several exams are performed daily and that there are few radiologists, in relation to the demand for specialized professionals, an automated screening would be of great help to identify possible regions containing tumors.

In several tasks, classification of medical images based on *Artificial Intelligence* (AI) presents excellent performance, sometimes with similar or superior quality to that of human experts (in [4], clinical task performance with and without AI aid systems are compared and discussed). Currently, there has been a great interest in the use of *Convolutional Neural Networks* (CNNs) for the analysis of medical images [5]–[9]. The traditional approach for classifying medical images is to extract characteristics using pre-defined filters. An example of the traditional approach is the method for classifying malignant and benign vertebral compression fractures in magnetic resonance images by using machine learning models in [10]. The vertebral bodies are manually segmented, and predefined shape and texture features are extracted and then used as inputs of the classifiers. An alternative approach was adopted in [9], where CNNs are used to classify the vertebral bodies. In this case, the segmented images are directly presented as inputs of the CNN. The great advantage of CNNs is that they are able to automatically extract interesting characteristics for classifying a given dataset [11]. As a consequence, it is not necessary to used pre-defined filters.

Shen et. al. [12] proposed to classify benign tumor and osteosarcoma by using both plain X-ray image features and metabolomic data. The features extracted from the images and the metabolomic data are classified by random forests or support vector machines. A similar approach was proposed in [13], but using plain X-ray image features and RNA-seq data. There is need to segment images in both approaches. In [14], deep learning is used, but for detecting osteosarcoma

---

This work was partially supported by São Paulo Research Foundation (FAPESP), under grants 2013/07375-0, 2019/07665-4 and 2020/01023-8, and National Council for Scientific and Technological Development (CNPq), under grants 305755/2018-8 and 118415/2019-0.

from histological images, and not from radiographs.

We propose a computer-aided diagnosis system based on CNNs for the classification of osteosarcoma on radiographs (plain X-ray images). The CNN should classify bone radiographs detecting the presence of osteosarcoma. The system should also indicate regions of the image that may contain the tumor. In order to indicate these regions on the image, we propose to split the image in windows and individually classify them by using the CNN.

In order to automatically generate the image windows and use them in training and testing the CNN, techniques for pre-processing, such as window exclusion and labeling, are proposed. By doing this, an advantage of the proposed method is that no manual pre-processing steps, e.g., segmentation and extraction of features, are necessary. Two CNNs are compared in the proposed system: i) a CNN trained from scratch; ii) a pre-trained CNN (VGG16). The traditional approach for classifying images is to extract features using predefined filters (features) and to use them as inputs of machine learning models. The CNNs are compared to four traditional machine learning models that use features extracted from the image windows as inputs: i) *multilayer perceptron* (MLP); ii) decision tree; iii) random forest; iv) MLP with feature selection.

The proposed computer-aided system based on CNNs is presented in Section II. In Section III, the results of experiments are presented. Finally, in Section IV, the paper is concluded.

II. MATERIALS AND METHODS

In the proposed system, a CNN classifies windows of radiographs containing bones. The CNN classifies the windows into one of two classes: normal and tumor (osteosarcoma). Next (Section II-A), the dataset used in the experiments is presented. The pre-processing procedures are also described. The system based on CNN is described in Section II-B, while the system based on machine learning models using predefined features is described in Section II-C.

A. Dataset and Pre-processing

The images dataset is composed of anonymized bone radiographs of patients diagnosed with osteosarcoma. The dataset was originally employed in a PhD research of one of the authors of this study at University of São Paulo. The use of the dataset was previously approved by the Ethics and Research Committee of the university. The original images were obtained using Computerized Radiography (CR) and were in the *Digital Imaging and Communications in Medicine* (DICOM) format. Each image was converted to the Portable Network Graphic (PNG) format. Figure 1 shows an example of the image in the PNG format.

A methodology for generating the inputs of the CNN similar to that adopted in [15] is used, in which each image is divided into small rectangular windows. Here, the objective is to classify each of the image windows into two classes: normal and osteosarcoma. In [15], the objective was to classify Focal Cortical Dysplasia in windows of brain images obtained by Magnetic Resonance Imaging.

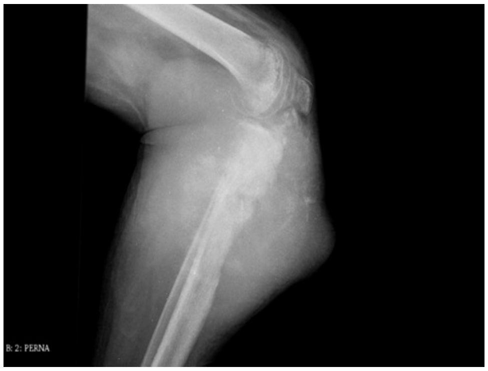


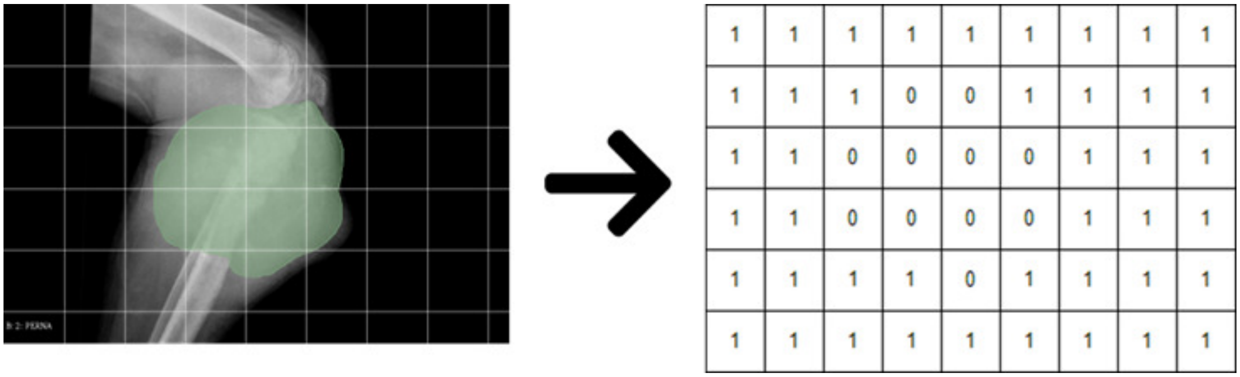
Fig. 1. Example of radiography in the PNG format.

Procedures were developed for automatically creating and labeling the image windows. In the windowing procedure, the radiography is cut into smaller square segments (windows), which are here used for training and testing the CNN. Figure 2 shows an example of the windowing procedure applied to the radiography shown in Figure 1.

Experiments (not shown here) were performed in order to compare the performance of the system for 50×50 pixels and 100×100 pixels windows. Best results were obtained for 100×100 windows; this size is used in the experiments presented in Section III. The 50×50 windows resulted in incorrectly classifying all examples with osteosarcoma, i.e., the model was unable to learn the relevant characteristics for the classification of tumors. Despite resulting in more examples for training, the 50×50 windows have a smaller amount of information relevant for the classification. This can be seen in the example presented in Figure 3, where it is easier to identify relevant characteristics in the window with 100×100 pixels.

In the labeling procedure, the windows used for training and testing the classifier are automatically labeled into one of two classes. The labels indicate the presence (0) or absence (1) of a tumor in the window. In order to create the dataset for training and testing the CNN, a radiologist manually marked the regions of the radiographs with tumor (osteosarcoma). These regions are represented in green in Figure 2. Using the image marked by the radiologist and the windowing procedure, the labeling procedure checks the percentage of the green color in the window. If the percentage is higher than a threshold, the window is marked with 0; otherwise it is marked with 1 (Figure 2). The threshold is equal to 19% in the experiments presented in Section III. This threshold was obtained in initial experiments (not shown here), where it was observed that thresholds with higher values resulted, for some radiographs, in the labeling of any window with tumor. The threshold equals to 19% ensures that at least one window is labeled with tumor on each of the images of the dataset.

The next procedure automatically eliminates windows with no relevant information, i.e., those windows that do not contain parts of the patient's body. The exclusion procedure counts all pixels with color in the range related to black and dark gray. If the window contains 95% or more black or dark gray



1	1	1	1	1	1	1	1	1
1	1	1	0	0	1	1	1	1
1	1	0	0	0	0	1	1	1
1	1	0	0	0	0	1	1	1
1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	1	1	1

Fig. 2. Example of the application of the windowing procedure (left). The figure also shows the mask (in green) created by the radiologist and the labels (right) automatically generated.

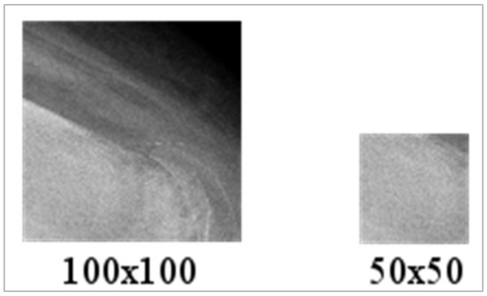


Fig. 3. Windows with 50×50 pixels and 100×100 pixels.

pixels (those with values between 0 and 5 on the luminance scale), the window is eliminated. Using a threshold smaller than 100% is necessary due to the presence of eventual noise in the images. Figure 4 shows an example of the exclusion procedure applied to the radiography shown in Figure 1.

B. CNNs

Python libraries and routines, such as TensorFlow [16] and Keras [17], were used in this work for implementing the CNNs. The open source library Pillow was used for image manipulation [18] and Google Colab [19] for running the CNNs. Google Colab is a free cloud service that offers free access to GPUs and easy sharing of codes. We propose two approaches for generating the CNN in the computer-aided diagnosis system.

1) CNN trained from scratch: The input of the CNN is the 100×100 pixels images and one single output indicates the presence or absence of osteosarcoma in the window. Experiments (not shown here) with a single radiography were carried out to select the architecture and hiper-parameters of the CNN trained from scratch. Accuracy was used to evaluate CNNs with different hiper-paramenters and architecture. The model with best results has five convolutional layers with 3×3 windows. The first two convolutional layers have 128 filters each, while the third and fourth layers have 64 filters each and the fifth layer has 32 filters. After the second, fourth and fifth convolutional layers, one MaxPooling layer is applied. MaxPooling layers have 2×2 windows. Finally, three fully connected layers, with respectively 8, 4 and 1 neurons, are added. In all convolutional and dense layers, the ReLu activation function is used, except in the last dense layer, where sigmoid function is used. Batch normalization is applied. The Adam optimizer with default parameters for TensorFlow is employed for adjusting the weights of the CNN.

2) Pre-trained CNN: Instead of creating and training a CNN from scratch, the pre-trained model employs a CNN with pre-defined architecture and pre-trained using a huge set of images. By doing this, features relevant for classifying a large number of types of images are discovered during training. These features are represented by convolutional and pooling layers. Here, the pre-trained CNN is *VGG16*, proposed by [20]. VGG16 obtained very good performance on the 2014 ImageNet Competition; it obtained 92.7% top-5 test accuracy on a dataset with more than 14 million images belonging to 1000 classes. An advantage of using pre-trained CNNs is that large architectures can be employed because we do not need to train the artificial neural network from scratch. VGG16 has many convolutional and pooling layers, and about 138 million parameters. Here, the VGG16 is combined with a fully connected hidden layer of 32 neurons, with ReLU activation function, and a last layer with 1 neuron for classification, with sigmoid activation function. The CNN was re-trained with the bone radiographs dataset by using default parameters for TensorFlow.

C. Machine Learning Models using Predefined Features

As an alternative to using CNNs, we can use traditional machine learning classifiers with inputs provided with predefined radiomic features. The radiomic features are extracted from the windows by using the *pyRadiomics* Library [21], that is an open source Python library for extracting radiomic features from medical images [21]. Here, *pyRadiomics* is applied to each 100×100 pixels image and the extracted features are used as inputs for the classifier. All features extracted from *pyRadiomics* are used, with the exception of 3D shape-based features. The features are:

- First Order Statistics: 19 features;
- Gray Level Co-occurence Matrix: 23 features;

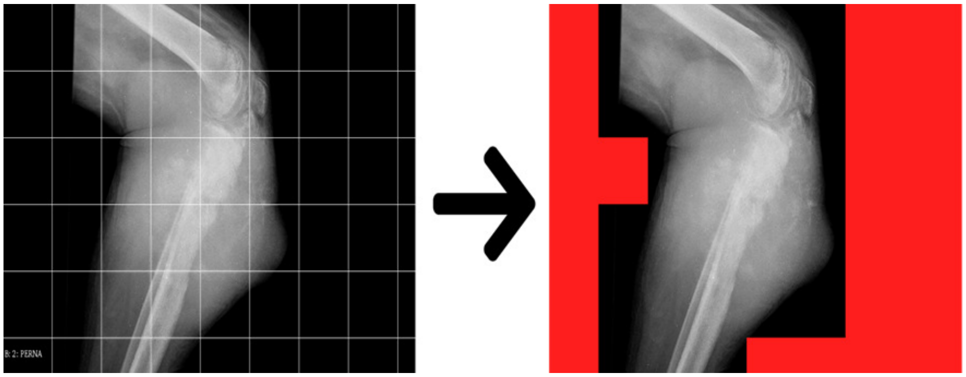


Fig. 4. Example of the application of the procedure for automatically eliminating windows with no relevant information (in red).

- Gray Level Size Zone Matrix: 16 features;
- Gray Level Run Length Matrix: 16 features;
- Neighboring Gray Tone Difference Matrix: 4 features;
- Gray Level Dependence Matrix: 14 features.

The classifiers are: decision tree, MLP, MLP with feature selection, and random forest. In the MLP with feature selection, the features extracted by the decision tree are used as inputs of the MLP. All models were implemented with default parameters of the Scikit-Learn Library [22].

III. EXPERIMENTS

The CNNs (Section II-B) are compared to four machine learning models that use features extracted from the image windows as inputs (Section II-C). All experiments were performed on a computer with 6GB of RAM and an Intel Core i5-4200U 1.6GHz processor.

A. Experimental Design

Radiographs of 45 patients were labeled by the radiologist, resulting in 2448 windows of 100×100 pixels. The exclusion procedure eliminated 1407 windows, resulting in a dataset with 1041 examples for training and testing the CNN. Ten-fold cross validation is used to evaluate the classifiers. Cross-validation is applied considering the division by patients and not by windows. In other words, the classifiers are trained using the windows of a subset of patients and tested using the windows of another subset of patients. This is done so that it is possible to observe the performance of each model for all the windows of a radiography, in the same way that it is done in a real-world situation. In addition, some windows on a radiography are expected to be similar; dividing the subsets by patient does not result in bias, that could be generated if windows of the same patient are used for training and testing the classifiers. There are 189 windows with tumor and 852 without tumor. In order to balance the dataset for training the models, the same number of windows (189) is used for each class during training. However, all 1041 examples of the dataset are used for testing the classifiers in cross-validation.

B. Results

The accuracy for the CNN trained from scratch was 74%. Table III-B shows the confusion matrix, while Figure 5 shows the classification results for the 6 first radiographs of the dataset. Figure 6 shows the ROC curve; the area under the curve (AUC) was 0.7307.

TABLE I
CONFUSION MATRIX OBTAINED BY THE CNN TRAINED FROM SCRATCH.

		Predicted Class	
		Tumor	Normal
Real Class	Tumor	134	55
	Normal	211	641

Table III-B shows the confusion matrix for the pre-trained CNN. The accuracy for the pre-trained CNN was 77%. The results of the pre-trained CNN for both classes were better than the results of the CNN trained from scratch. The pre-trained CNN uses a larger and more complex architecture than the CNN trained from scratch. This architecture showed to be more effective for this image classification problem. In addition, the pre-trained CNN was previously trained with a large image dataset that allowed to discover features useful for the classification of different types of images. Those features were useful for classifying the dataset used in the experiments. Better results could be obtained by the CNN trained from scratch if a larger dataset were employed.

TABLE II
CONFUSION MATRIX OBTAINED BY THE PRE-TRAINED CNN.

		Predicted Class	
		Tumor	Normal
Real Class	Tumor	159	30
	Normal	209	643

The accuracy, sensitivity, and specificity for all models are presented in Table III-B. The pre-trained CNN obtained the best performance among all models. It obtained the better accuracy and sensitivity (tied with the MLP with feature selection), and the second better specificity (0.75 against 0.76 of the MLP).

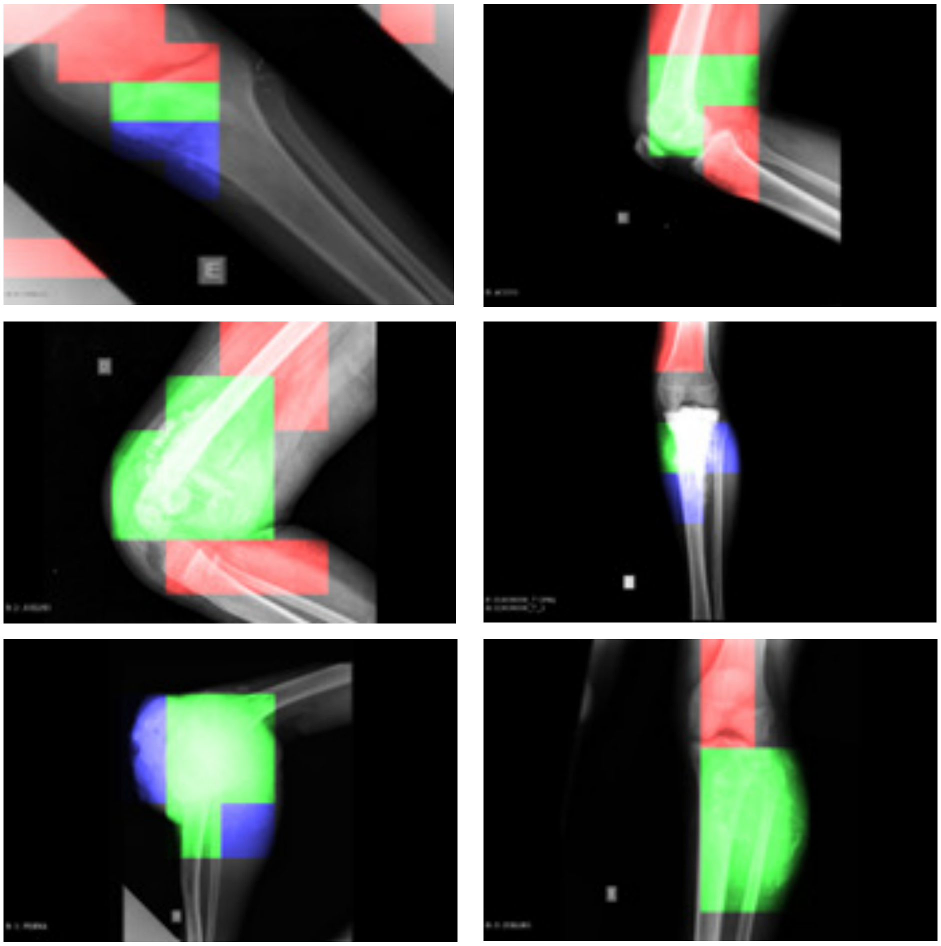


Fig. 5. Classification results of the CNN trained from scratch for the 6 first radiographs of the dataset. The windows in green represent true-positive examples (windows with tumor that are correctly classified). The windows in red represent false-positive examples (windows without tumor that are classified as with tumor). The windows in blue are false-negative examples (windows with tumor that are classified as with tumor). The non-colored windows are true-negative (windows without tumor that are correctly classified).

TABLE III

TEN-FOLD CROSS VALIDATION RESULTS FOR ALL MODELS. THE BEST RESULTS ARE IN BOLD.

	Decision Tree	Random Forest	MLP	MLP with Feature Selection	CNN trained from scratch	Pre-trained CNN
Accuracy	0.69	0.71	0.76	0.71	0.74	**0.77**
Sensitivity	0.66	0.79	0.75	**0.84**	0.71	**0.84**
Specificity	0.69	0.69	**0.76**	0.69	0.75	0.75

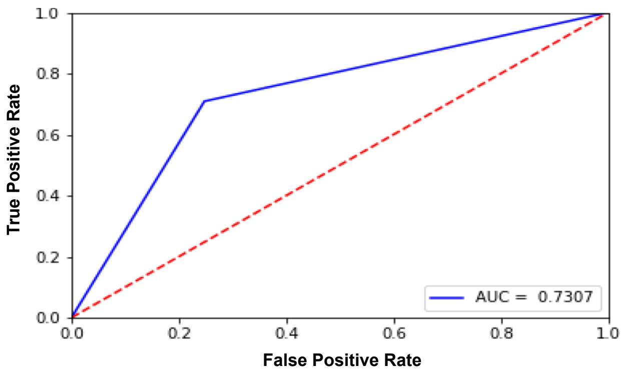


Fig. 6. ROC curve for the CNN trained from scratch.

The CNNs employed here have one output with sigmoid activation function. The criterion adopted here for classifying the windows is that a tumor is detected if the output is smaller than 0.5, and the class is normal otherwise. Alternatively, the radiologist can analyze the value of the output as a confidence indicator for the presence of tumor. An example is presented in Figure 7; figures like this can be automatically generated, helping the radiologist when making a decision.

IV. CONCLUSIONS

A computer-aided diagnosis system based on CNNs for the classification of osteosarcoma on radiographs is proposed. In order to indicate the regions with tumor, the image is divided in windows. These windows are individually classified by using

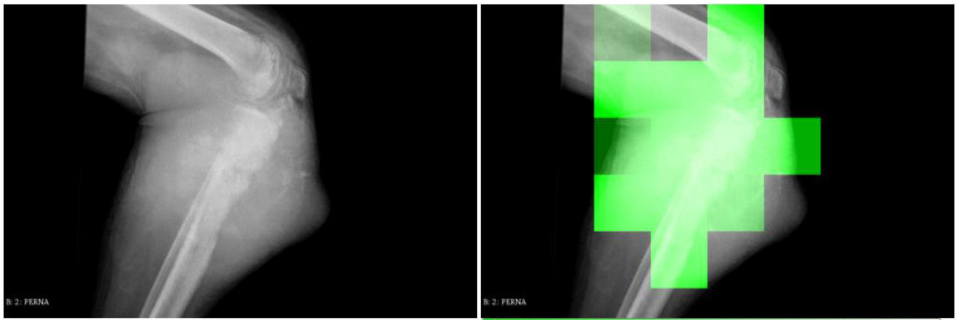


Fig. 7. Classification generated by the CNN (right) for the radiography shown on the left. Outputs smaller than 0.5 are colored in green. The stronger the intensity of the green, the smaller the CNN output. The output of the CNN can be used as a confidence indicator for the presence of tumor.

the CNN. The output of the CNN can also be used as a confidence indicator for the presence of tumor, helping the radiologist when making a decision.

The main attraction of the proposed methodology is that, after training, all pre-processing steps are automatic, i.e., the radiologist does not need to segment the images, extract features or perform any manual pre-processing steps. Procedures were proposed here for automatically creating the windows, excluding irrelevant windows, and labeling the examples for training and testing the models. The best results were obtained for windows with 100×100 pixels, but the classification system can be used with different windows size.

When compared to the CNN trained from scratch and to 4 machine learning models that use pre-defined features as inputs, the best performance was obtained by the pre-trained CNN. The accuracy obtained by the pre-trained CNN was 0.77, while the sensitivity and specificity were respectively 0.84 and 0.75. For future work, it would be interesting to increase the dataset with more radiographs, including those with fractured bones. In addition, an overlapping strategy for generating the windows can be investigated.

REFERENCES

[1] E. Onikul, B. Fletcher, D. Parham, and G. Chen, "Accuracy of mr imaging for estimating intraosseous extent of osteosarcoma." *AJR. American Journal of Roentgenology*, vol. 167, no. 5, pp. 1211–1215, 1996.

[2] D. D. Moore and H. H. Luu, "Osteosarcoma," in *Orthopaedic Oncology - Primary and Metastatic Tumors of the Skeletal System*, T. Peabody and S. Attar, Eds. Springer, 2014, vol. 162, pp. 65–92.

[3] Z. S. Kundu, "Classification, imaging, biopsy and staging of osteosarcoma," *Indian Journal of Orthopaedics*, vol. 48, pp. 238–246, 2014.

[4] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi, "Do as ai say: susceptibility in deployment of clinical decision-aids," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–8, 2021.

[5] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[6] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2017.

[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[9] R. Del-Lama, R. Candido, N. Chiari-Correia, M. Nogueira-Barbosa, P. Azevedo-Marques, and R. Tinós, "Computer-aided diagnosis of vertebral compression fractures using convolutional neural networks and radiomics," *Submitted to Journal of Digital Imaging*, 2021.

[10] L. Frighetto-Pereira, R. M. Rangayyan, G. A. Metzner, P. M. Azevedo-Marques, and M. H. Nogueira-Barbosa, "Shape, texture and statistical features for classification of benign and malignant vertebral compression fractures in magnetic resonance images," *Computers in Biology and Medicine*, vol. 73, pp. 147–156, 2016.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] R. Shen, Z. Li, L. Zhang, Y. Hua, M. Mao, Z. Li, Z. Cai, Y. Qiu, J. Gryak, and K. Najarian, "Osteosarcoma patients classification using plain x-rays and metabolomic data," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 690–693.

[13] O. Alge, L. Lu, Z. Li, Y. Hua, J. Gryak, and K. Najarian, "Automated classification of osteosarcoma and benign tumors using rna-seq and plain x-ray," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1165–1168.

[14] D. Anisuzzaman, H. Barzekar, L. Tong, J. Luo, and Z. Yu, "A deep learning study on osteosarcoma detection from histological images," *arXiv preprint arXiv:2011.01177*, 2020.

[15] S. Silva, F. Simozo, L. M. Junior, and R. Tinós, "Uso de redes neurais convolucionais para identificar displasia cortical focal em pacientes com epilepsia refratária," in *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, 2020.

[16] M. Abadi, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[17] N. Ketkar, "Introduction to keras," in *Deep learning with Python*. Springer, 2017, pp. 97–111.

[18] A. Clark, "Pillow (pil fork) documentation: Release 6.2.0.dev0," https://www.realmoon.net/wordpress/wp-content/uploads/2019/07/pillow.pdf, 2015.

[19] E. Bisong, "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 2019, pp. 59–64.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] J. J. Van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[22] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.