International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Transfer Learning and Majority Voting Approaches for Osteoporosis Classification

Mohamad Melad Ashames[1], Murat Ceylan[2], Rachid Jennane[3]

Submitted: 23/10/2021 Accepted : 01/12/2021

Abstract: Osteoporosis is a systemic skeletal disease characterized by low bone mass density and deterioration of the micro-architectural structure of the bone tissue, increasing bone fragility, and the probability of fracture. In this study, we propose a non-invasive method for osteoporosis classification using X-ray images (plain radiographs) of the ankle. Convolutional Neural Networks along with Data Augmentation techniques and Deep Transfer Learning Architectures are combined to classify X-ray images of healthy and osteoporotic patients. The proposed approach achieved an accuracy of 99% using ResNet50, and 100% with GoogleNet.

Keywords: *CNN, data augmentation, transfer learning, osteoporosis, X-ray*

1. Introduction

Osteoporosis is a disease in which the mineral density of bones is reduced, causing porosity and fragility of the bones, hence the risk of fracture is greatly increased [1-2]. The loss of bone occurs silently and progressively, thereby it is occasionally referred to as a silent disease. Often there are no symptoms until the first fracture occurs. The most common fractures associated with osteoporosis occur in the hip, spine, and wrist, and the likelihood of hip and spine fractures increases with age in women and men [1-2]. The disease in the preclinical period is characterized by low Bone Mineral Density (BMD) without fractures. Osteopenia refers to low bone density which is not lower enough to be considered as osteoporotic. People with low BMD may become osteoporotic in the future.

It is estimated that there are more than 200 million people with osteoporosis in the World [3]. There is a risk of osteoporotic fractures in every 3 women and every 5 men [4]. The diagnosis of osteoporosis is based on the values obtained using Dual Energy X-Ray Absorptiometry (DEXA) exam and the presence of a fracture. DEXA is the most widely used technique to measure Bone Mineral Density (BMD) and recommended by the National Osteoporosis Foundation (NOF) for the diagnosis of osteoporosis [5]. Using the T-Score, BMD values are evaluated ($-1 \leq$ T-score ≤ -2.5 means low BMD or osteopenia; T-score ≤ -2.5 indicates the presence of osteoporosis).

Despite DEXA's efficiency for evaluating BMD, it has some disadvantages as low availability, high-price, and over-size of the device. Also, osteophyte formations around the joints caused by the disease (especially the knee, hip, and vertebral column area), sometimes cause false-positive DEXA results. This results in patients who are osteoporotic while not receiving adequate treatment. On the other hand, bone X-ray images cannot show the symptoms since the bone microarchitecture changes due to osteoporosis are visible on X-rays only in case of severe osteoporosis. Thus, depriving the patient of preventive measures and treatments enabling to avoid the aggravation of the disease. Computed Tomography (CT), as well, does not show any symptom predicting the disease. In addition, there are physics-based models employing Finite Element Analysis (FEA), which have shown impressive performance to evaluate the bone microarchitecture stiffness non-invasively. However, these models have high computational cost, making them incommodious for clinical use.

Several studies in the literature proposed different techniques to predict osteoporosis. Jennane et al. estimated the 3D similarity parameters from 2D trabecular bone images [6]. In [7], a classification of arthritic and osteoporotic bone samples was performed with an adaptive Neural Fuzzy Interference System (ANFIS), Support Vector Machines (SVM) and Genetic Algorithms (GA). Only 18 images were used in the study. In [8], authors used image processing and GA. The same 18 images were used in this study as well. In [9], Houam et al. performed trabecular bone tissue classification using Wavelet Coefficients and one-dimensional local binary patterns in high-pass bands. The K-Nearest Neighbor classifier was used and the value of the Area Under the Curve (AUC) was a maximum of 0.85.

Another approach applied the Wavelet transform to extract features from CT images, while Artificial Neural Networks (ANN) and SVM were used for the classification task [10]. In [11], authors tested the effectiveness of a Multilayer Perceptron (MLP) in discriminating between osteoporotic and control cases. They used k-fold Cross-Validation (CV) to increase the model's accuracy and reliability. Here, 120 X-ray images of the calcaneus bone were used.

In [12], the bone fragility was evaluated by combining BMD and texture analysis. Obtained results showed an AUC = 82% at most. The classification accuracy was around 70%. In [13], Jennane et al. characterized osteoporosis using fractal analysis on X-Ray images. The fractional Brownian motion (fBm) model was used to

[1]*Electric Electronic Eng., Eskisehir Osmangazi University, Eskisehir-26000, Turkey. ORCID ID: 0000-0002-2837-1343*
[2]*Electric Electronic Eng., Konya Technical University, Konya-42000, Turkey. ORCID ID: 0000-0001-6503-9668*
[3] *University of Orleans, IDP Laboratory, UMR CNRS 7013, Orléans-45000, France. ORCID ID: 0000-0002-8032-8035*
** Corresponding Author Email: muhashames@gmail.com*

extract features from 77 X-ray images of the calcaneus and the SVM was used as a classifier. In [14], Nasser et al. used a Stacked Sparse Autoencoder (SSAE) to extract features along with the SVM classifier to discriminate between two groups of osteoporotic and control patients. Ciusdel et al. trained a CNN model on a large database of synthetically generated cancellous bone anatomies [15]. The performance of the trained model was assessed by comparing the predictions against a FEA model computed on a separate test data set.

In [16], Tomita et al. developed a system for detecting osteoporotic vertebral fractures in which a deep CNN was used to extract radiological features from each slice of 1432 CT scans. A Long Short-Term Memory (LSTM) network processed the extracted features to make the final diagnosis for a full CT scan. The CNN-LSTM model achieved an accuracy of 89.2%.

In [17], Selaa and Pulunganb extracted trabecular area present on digital dental radiographic images to identify osteoporosis. The Multilayer Perceptron was used to predict the presence of osteoporosis using statistical texture analysis.

In [18], authors introduced a classification decision strategy based on maximum a posteriori probability and an approach for adjusting the classification decision criteria to discriminate between two groups of osteoporotic and control patients. Su et al. [19] combined deep Convolutional Neural Network (CNN) and several hand-crafted features to classify osteoporotic and control subjects.

Transfer Learning (TL) was also applied to detect osteoporosis using X-ray images. In [20], the TL model VGG-16 was used to detect low bone density in Dental Panoramic Radiograph (DPR) images, it was shown that a fine-tuned pre-trained VGG-16 model can reach an accuracy of 0.84.

In this study, using X-ray images of the calcaneus bone, our aim is to classify the patients of two populations composed of osteoporotic and control subject. The proposed approach combines both Artificial Intelligence (AI) and image processing techniques. Our proposed approach is based on the use of CNNs along with transfer learning methods. Different data augmentation methods are also used to increase the number of samples and to extract discriminatory features.

We investigated the impact of sharpness, contrast, and brightness adjustments on deep networks' generalizability and their capability to improve osteoporotic-control data classification accuracy. Furthermore, we compared the performance of six different CNN based sequential and residual networks. The results suggest that our proposed transfer learning-data augmentation approach can play a fundamental role in the development of a novel and effective method to support early diagnosis of osteoporosis, which can aid earlier interventions and prevent further disease progression.

The rest of the paper is organized as follows. Section 2 is divided into 4 subsections. The first subsection presents the data. The second subsection describes the proposed methodology to augment the number of samples. The third subsection covers the transfer learning and the training process. Section 3 details the experimental results. Finally, a conclusion is presented in Section 4.

2. Material and Methods

2.1. Data

The database consists of X-ray images collected from 174 women aged between 40 and 92 years. The patients were hospitalized at Orleans Hospital between November 2004 and February 2006. Because age has an influence on bone density and on trabecular bone texture, the fracture cases were age-matched with the control cases. Also, all fracture cases were reviewed by experienced investigators who considered the diagnosis of fragility fracture if it occurred after the age of 40 years. The cases were described as either spontaneous fractures, fractures resulting from strenuous activity, fractures after falls from standing height or less (low trauma energy) and following radiologic data. The selected database contains images labelled as 87 osteoporotics (OP) and 87 controls (CN). Among OP patients, there were 21 patients diagnosed with hip fractures, 23 patients with wrist fractures, and 22 patients with vertebral fractures. The remaining 21 patients had different fractures [21].

X-ray acquisition of the heel enabled the selection of a similar Region Of Interest (ROI) for each subject by identifying anatomical landmarks as described in [21]. These anatomical landmarks were localized on each image by an experienced operator, allowing positioning of the ROI (1.6 × 1.6 cm2) performed by a software device (Fig. 1). The size of each image is 400x400 pixels.

The heel bone was selected because it is surrounded by limited soft tissues that may increase the instability of the model to be used. Figure 1 shows the calcaneus bone and the selected ROI.

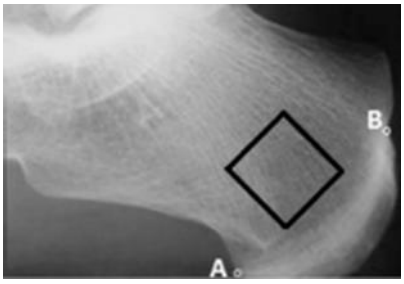

Fig. 1. The calcaneus bone and the selected ROI

The bone microarchitecture of osteoporotic and healthy bone radiograph images is visually similar, making the classification task very challenging. An example of an OP and a CN image is shown in Figure 2.

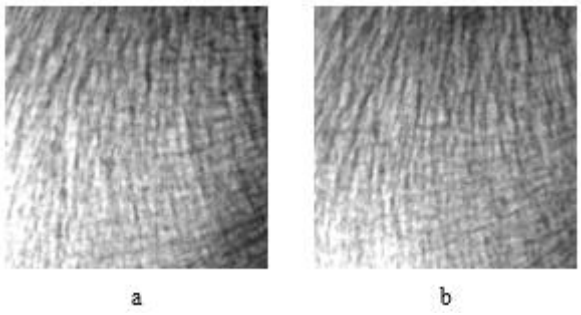

Fig 2. Two representative images from the database. (a) OP image; (b) CN image

Data-driven models that are trainable to learn relevant features from the raw input are becoming common, especially with the concept of feature learning, which is the very strength of Deep Learning (DL) [22].

In the next sections, a comparison between different TL models is presented. As large dataset is crucial for the performance of DL models, the performance of the model can be improved by augmenting the available original images. This can also be done through pre-processing of the images. Figure 3 presents the block diagram of the proposed approach.

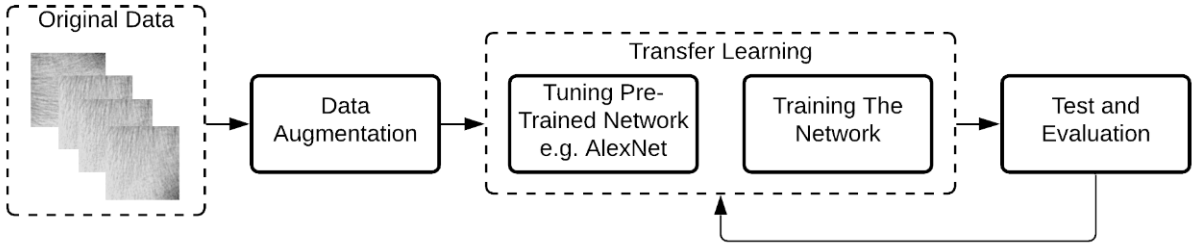

Fig 3. Block diagram of the proposed approach for the classification of OP and CN data

2.2. Data Augmentation

To properly train a DL model it needs to have high generalizability. Generalizability refers to the variety of training data and the ability of the model to correctly classify blind test data. To make the model more general, it needs to cover several data possibilities such as different orientations or color scales of the samples. Sometimes, a DL model would misclassify an image because of a different version of the image when it has a different color range, orientation or position, which the training data didn't include such a version. Also, models with poor generalizability are often overfitting the training data. Which means that they reach high training accuracy while the validation accuracy is low. To create an efficient DL model, the validation accuracy needs to continue increasing with the training accuracy, and the validation error must continue decreasing with the training error. Data augmentation [23] helps building a more general model that covers as many possibilities as it can, and it is one of the most useful techniques that prevents the model from overfitting.

In this study different conventional techniques were used to augment the original dataset (174 images). Data augmentation procedures were applied through three steps to obtain three different data versions to be used. The aim is to check the impact of data augmentation and the relation between the size of the input images and the resulting classification accuracy. Figure 4 details the different steps of the data augmentation procedure.

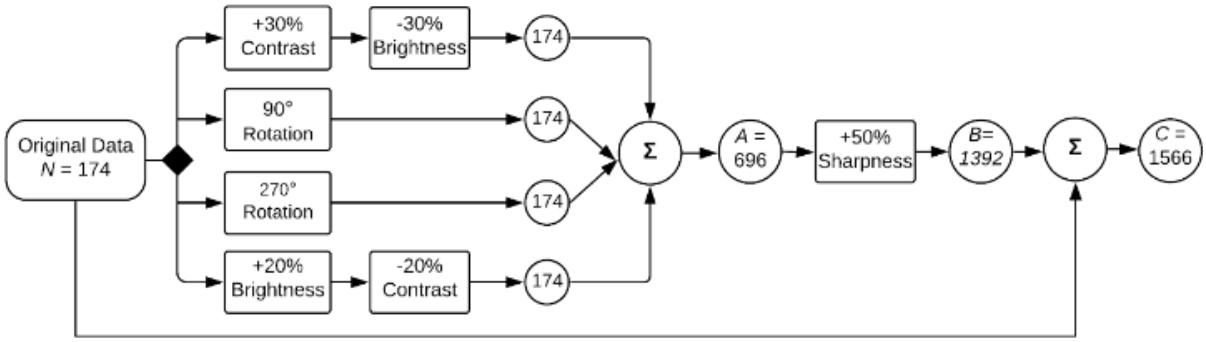

Fig 4. Blockdiagrm of the data augmentation procedure

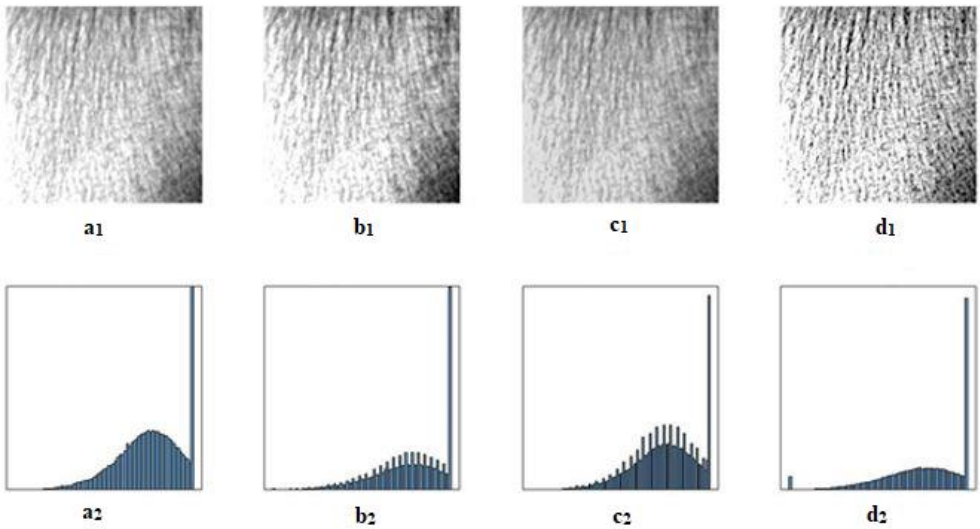

Fig 5. A CN image and it's adjusted versions with the histogram of each image. Original CN image (a_1, a_2); 30% contrast increase applied (b_1, b_2); 30% brightness decrease applied (c_1, c_2); 50% sharpness increase applied (d_1, d_2)

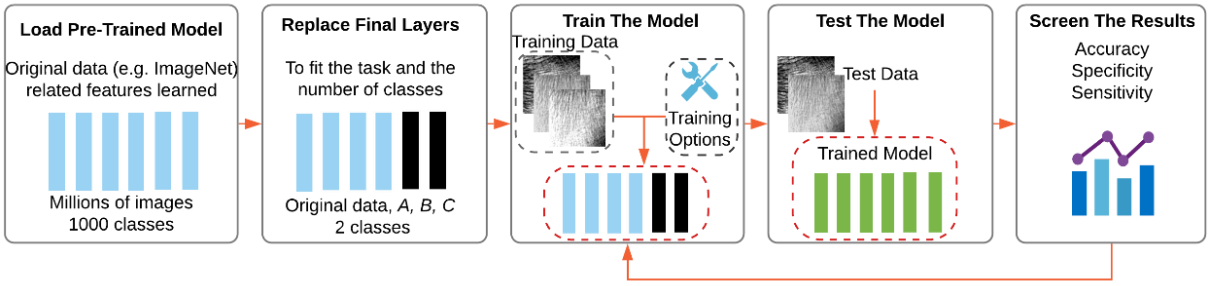

Fig 6. Procedure of using a pre-trained model (Transfer Learning)

At the first step, the N original images were rotated once 90 degrees and another time 270 degrees. The rotation was used both to increase the data size and the generalizability of the model. At the same step, the contrast of each image was increased by a ratio of 30% while the brightness was decreased with a ratio of 30%. Also, 20% brightness increasing, and 20 % contrast decreasing were performed on each image to fill the gap with the opposite situation with more brightness and less contrast. At this step, the dataset, A, counted 696 images. Contrast and brightness increase and decrease were applied to extract more discriminatory information from the original images. Contrast adjustment remaps image intensity values to help extracting more information. An image with a good contrast has sharp differences between black and white. Contrast adjustment was executed using histogram equalization technique. Such pixel-domain enhancement techniques are preferable due to their low computational cost and simple parameter setting comparing to transform-domain contrast enhancement techniques [24].

At the second step, all the 696 images were sharpened by a ratio of 50%, providing a new 696 images, and a database B composed of 1392 images. Sharpening was implemented with a radius of 3. The goal of using such technique is to highlight the trabeculae and the features in the bone structure. Sharpening images increases the contrast along the edges where different colors meet [25]. Figure 5 presents a CN image and its adjusted versions with the histogram of each image. Histograms in Figure 5 are only added to show the differences between the image versions.

Finally, the original data, N (174), were added to the resulting data, B (1392), which provided a database C containing 1566 images.

2.3. Deep Transfer Learning

Transfer Learning (TL) is a popular machine learning method, which enables building a model from a previously pre-trained model [26]. It is a popular approach in deep learning where pre-trained models are used as the starting point. A pre-trained model is a model trained on a large dataset (e.g., ImageNet) to solve a specific problem.

At each layer of the Deep Neural Network (DNN), data related features are learned, each layer is connected to a deeper layer via a set of trainable weights, thus the previously learned features represent input data for the next deeper layer. Input data and learned weights are convolved to calculate a new feature map, and the results are forwarded to an activation function (see section 2.2.3) [26].

These DNN usually include convolutional and pooling layers. Convolutional layers are the responsible of the previously mentioned feature extraction task. In general, convolution is the implementation of a sliding window function on an image's matrix. The window function here is often referred to as a kernel or filter.

Meanwhile, pooling layers comes after convolutional layers and they are responsible for performing sub-sampling on the given input matrix, summarizing it by applying different filters. Max Pooling takes the maximum value of the filter, Min Pooling takes the minimum value of the filter and Average Pooling computes the average value of the filter. Thus, reducing the spatial resolution of the feature maps. Also, pooling layers play an important role in reducing model overfitting.

Convolutional and pooling layers are followed by fully connected layers, which are responsible for the classification task. In the classification task, fully connected layers use the Softmax function as a default, while sometimes it's replaced with a Support Vector Machine (SVM). In this study, the Softmax function is used for the classification of the data. The Softmax function produces an output value between 0 and 1. Every value represents the probability of belonging to one class and the sum of all the probabilities is equal to 1 [27].

2.3.1. Tuning the Pre-Trained Models

To use a pre-trained model in any classification problem, tuning according to the problem is needed. There are different steps to do so, but all of them includes two important steps. The first one is changing the classifier used in the original model if it is not suitable for the given problem, and even if the same classifier must be used, one must adjust the number of classes according to the classification problem to be solved. The second step consists in resizing the input data to meet the requirements of the pre-trained model. Figure 6 illustrates the process of using a pre-trained model in a classification task. In this study different TL models were used to classify the various augmented versions of the data.

2.3.2. Using a Pre-Trained Model

Various pretrained models such as VGG [28], InceptionV3 [29], ResNet [30], AlexNet [31], GoogleNet [32], MobileNetV2 [33], DenseNet [34], etc. show great performance for classification tasks. In this study, five pretrained models (AlexNet, GoogleNet, ResNet50, MobileNetV2 and InceptionV3) were selected to classify OP and CN images. Each model has a different depth, number of parameters, and input size. But they are all trained on ImageNet (a large image database consisting of more than 14 million images) [35]. Fully connected layers in each architecture were modified for the binary classification of the two classes (OP and CN). The architectures were used for End-2-End training and no layers were frozen. ImageNet weights were also utilized for the previously mentioned architectures, so they were not trained from scratch. Also, when required, the size of the input images was resized to match the default size of each model. Table 1 presents the input size, depth, number of parameters for the models used in this study.

Table 1. Input size, depth, number of parameters for the pretrained models.

Model	Data Trained On	Input Size	Number of Layers	Number of Parameters
AlexNet	ImageNet	227x227	8	60,000,000
GoogleNet	ImageNet	224x224	22	6,797,700
Resnet50	ImageNet	224x224	50	25,636,712
MobileNetV2	ImageNet	224x224	53	3,538,984
InceptionV3	ImageNet	299x299	48	23,851,784

2.3.3. Training Transfer Learning Models

The training of the models was fulfilled in several experiments. In the first experiment 1 (E1), only the original dataset was used to train the models, while in the following experiments, E2, E3, E4 the training was achieved using datasets A, B, and C, respectively. The aim is to evaluate the influence of the carried-out data augmentation techniques on the model's accuracy. As for data partitioning, 80% of the datasets were partitioned as training data at each experiment. Images were shuffled and the mini-batch size value was set to 5. The number of epochs was held to 20 throughout the process. Hence, the number of iterations can be counted by the following equation:

$$Number\ of\ Iterations = (Number\ of\ Training\ Images \div Mini\ Bach\ Size) \times Number\ of\ Epochs \quad (1)$$

ReLU (Rectified Linear Unit) Activation Function, also known as the Ramp Function, was used to determine the output of the models. An activation function is a switch that decides whether a neuron should take a value of zero or one. ReLU activation function outputs the input directly if it is positive, and outputs zero if it is not.

It is common to use the Tanh (hyperbolic tangent) or the sigmoid functions as activation functions to train a learning model. But they are not as efficient as ReLU when used in deep networks, due to the vanishing gradient problem. Meanwhile, ReLU has been accomplishing much better performance, especially for image classification tasks, thanks to its ability to overcome the vanishing gradient problem by forcing the negative values to zero. Thus, enabling the models to learn faster. Another reason for the ReLU being faster than it's equivalent sigmoid or Tanh functions is the absence of exponentials, therefore less computational power is needed [27]. Figure 7 illustrates the three activation functions (Sigmoid, Tanh, and ReLU).

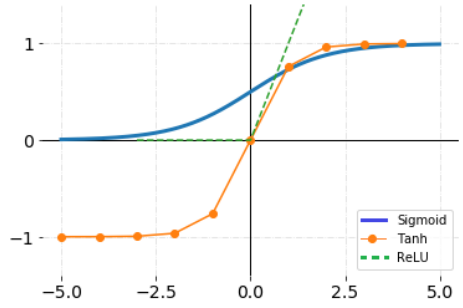

Fig 7. Sigmoid, Tanh, and ReLU activation functions

Krizhevsky et al. [31] showed that a deep neural network could be trained much faster using the ReLU function compared to when it is trained using the saturating activation functions like Tanh or sigmoid. Using ReLU, AlexNet achieves 25% training error rate six times faster than it's reached rate using Tanh. These findings encouraged us to use the ReLU function for our classification task. As shown in table 2, Stochastic Gradient Descent with Momentum (SGDM) is the optimizer retained in this work. An optimizer is responsible for reducing the losses of a learning model. To improve the performance of classification, the optimizer continuously updates the model according to the calculated loss function [36]. SGDM is an improved version of the classical optimization algorithm SGD (Stochastic Gradient Descent) in which momentum is added to accelerate gradients vectors in the optimal direction and avoid being stuck in a false local minimum point, making it notably more efficient than the classical SGD optimizer [37-38].

Other training options are shown in Table 2. All the options were held the same throughout the study.

Table 2. Training options throughout the study

Training Data	80%, 3-Fold Cross-Validation
Learning Rate	0.0001
Activation Function	ReLU
Optimizer	SGDM
Number of Epochs	20
Mini Bach Size	5

2.3.4. Performance Evaluation

The confusion matrix is used to summarize the performance of a classifier. Classification accuracy alone can be misleading if there is an unbalanced number of observations in each class or in case of more than two classes in the dataset. The confusion matrix gives a better idea about the classification of the model, emphasizing better the misclassified subjects in the correct class. The following metrics are used to compute the confusion matrix: True Positive (TP), which is the number of OP patients correctly identified, False Positive (FP), which is the number of CN subjects incorrectly identified, True Negative (TN), which is the number of CN subjects correctly identified, False Negative (FN), which is the number of OP patients incorrectly identified, Sensitivity (Sn), which tests the ability of the classifier to identify positive results and Specificity (Sp), which test the ability of the classifier to identify negative results. The Accuracy (Acc) of classification of the subjects is defined as:

$$\text{Acc} = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

Sn also called True Positive Rate (TPR) is given by:

$$Sn = TP/(TP + FN) \quad (3)$$

Sp also called True Negative Ratio (TNR) is given by:

$$Sp = TN/(TN + FP) \quad (4)$$

2.3.5. K-Fold Cross-Validation

As there is a need of lot of data to train a neural network, removing a part of it for validation poses a problem of underfitting. By reducing the training data, we risk losing important details from the images, which in turn increases error induced by bias. To double-check on the efficiency of the proposed approach and avoid the risk of losing discriminatory details, all the models were trained several times using the K-fold cross-validation.

K-fold cross-validation consists in dividing the set of data into K subsets, and the model is trained K times. Each time, one of the K subsets is used as a test set and the other K-1 subsets are used together to form a training set. Then the average error across all K trials is computed. The advantage of this method is that each sample is given the opportunity to be used in a test set once, and in a training set K-1 times. In this study, K-fold cross-validation was implemented only on Experiment 7 (section 3.7.) in which the dataset C was used. To obtain subsets of size integer, K value must be a factor of the number of data samples (in this case 1566). In this study, K was set to 3 and the dataset C was divided into 3 subsets.

3. Results & Discussion

As explained in Section 2, different datasets were used, which included ROIs extracted from X-ray images of the calcaneus bone taken from OP and CN subjects. Since there is a positive correlation between the number of samples and the accuracy of the model, data augmentation was used to enhance the learning of the model. Contrast, brightness, and sharpness modification were implemented with the ambition of enhancing X-ray scans and extracting unseen information that plays a role in the classification task. AlexNet, MobileNetV2, GoogleNet, InceptionV3, and Resnet50 architectures were used to classify the four datasets (original, A, B and C). The obtained results of the different experiments are shown in the next Sections. Effects of data augmentation to improve the accuracy of classification were evaluated through seven experiments described in the following subsections.

3.1. Experiment 1: Classification of the Original Data

This experiment aims to compare the classification result obtained using the N original data to those obtained through data augmentation procedure. The test data consisted in 34 images (17 images labelled as CN and 17 images labelled as OP), which represents a percentage of 20% of the whole original dataset. Obtained classification rates (Acc, Sn, Sp) using the different pretrained models (AlexNet, GoogleNet, ResNet50, MobileNetV2 and InceptionV3) are shown in Table 3. As can be seen, Acc values varied from 47% with InceptionV3 to 58.8% with ResNet50. As expected, all of the models couldn't accomplish promising results due to the lack of data samples. The number of both accurately classified and misclassified subjects (TP, TN, FP, FN) are given in Table 4. Looking to ResNet50's results (best performer of this experiment), 6 out of 17 CN labelled images were misclassified (FP), and 8 out of 17 OP labelled images were classified as CN (FN).

3.2. Experiment 2: Classification of Dataset A

Dataset A was realized by applying both 90 and 270 degrees rotation, followed by two other steps where images contrast and brightness were adjusted by different rates (Section 2.1). Dataset A contains 696 images half of which are OP, while the rest are CN. At this experiment, the test data included 140 images divided into 70 CN and 70 OP. Tables 3 and 4 present the evaluation metrics. As can be seen, the accuracy values obtained using the five models (AlexNet, GoogleNet, ResNet50, MobileNetV2 and InceptionV3) dramatically improved to vary from 77.8% with MobileNetV2 to 82.1% with InceptionV3. ResNet50 and GoogleNet reached the same Acc value of 80%, while AlexNet performed better with an Acc of 81.4%. As can be seen in Table 4, TN rates at all models are greater than their corresponding TP rates. Hence in this experiment all the models are more specific (Sp) than sensitive (Sn).

3.3. Experiment 3: Classification of Dataset B

The trabeculae of the bone structure in the X-ray images are blurry and quite pale. Therefore, the sharpness of images dataset A was intensified by 50%, making the images much more comprehensible. Adding the new sharp images to dataset A, a new dataset of 1392 images, B, was obtained. As can be seen in Table 3, Acc, Sp, and Sn values were impressively boosted to reach an Acc of 98.5% with GoogleNet and 97.4% with InceptionV3. AlexNet, ResNet50 and MobileNetV2 models reached Acc values of 95.3%, 96%, and 94.9, respectively. Table 4 shows the decrease of FN and FP rates. Note that GoogleNet succeeded to accurately classify all the CN labelled 139 test images.

Table 3. Obtained Acc, Sn, Sp values for each model with each dataset

Trial	Metric	AlexNet	GoogleNet	ResNet50	MobileNetV2	InceptionV3
Experiment 1	Acc	50%	52.9%	58.8%	58.6%	47%
174 images	Sp	35.2%	58.8%	52.9%	64.7%	64.7%
N	Sn	64.7%	47%	64.7%	52.9%	29.4%
Experiment 2	Acc	81.4%	80%	80%	77.8%	82.1%
A, 696 images	Sp	88.5%	81.4%	80%	84.2%	82.8%
	Sn	74.2%	78.5%	80%	71.4%	81.4%
Experiment 3	Acc	95.3%	98.5%	96%	94.9%	97.4%
B, 1392 images	Sp	93.5%	100%	96.4%	92.1%	99.2%
	Sn	97.1%	97.1%	95.6%	97.8%	95.6%
Experiment 4	Acc	98.4%	97.1%	99.02%	98%	98.4%
C, 1566 images	Sp	100%	96.1%	98%	97.4%	96.8%
	Sn	96.8%	98%	100%	98.7%	100%

Table 4. Obtained TP, TN, FP and FN values for each model with each dataset

Trial	Metric	AlexNet	GoogleNet	ResNet50	MobileNetV2	InceptionV3
	TP	11	8	11	9	5
Experiment 1	TN	6	10	9	11	11
N, 174 images	FP	6	9	6	8	12
	FN	11	7	8	6	6
	TP	52	55	56	50	57
Experiment 2	TN	62	57	56	59	58
A, 696 images	FP	18	15	14	20	13
	FN	8	13	14	11	12
	TP	135	135	133	136	133
Experiment 3	TN	130	139	134	128	138
B, 1392 images	FP	4	4	6	3	6
	FN	9	0	5	11	1
	TP	152	154	157	155	157
Experiment 4	TN	157	151	154	153	152
C, 1566 images	FP	5	3	0	2	0
	FN	0	6	3	4	5

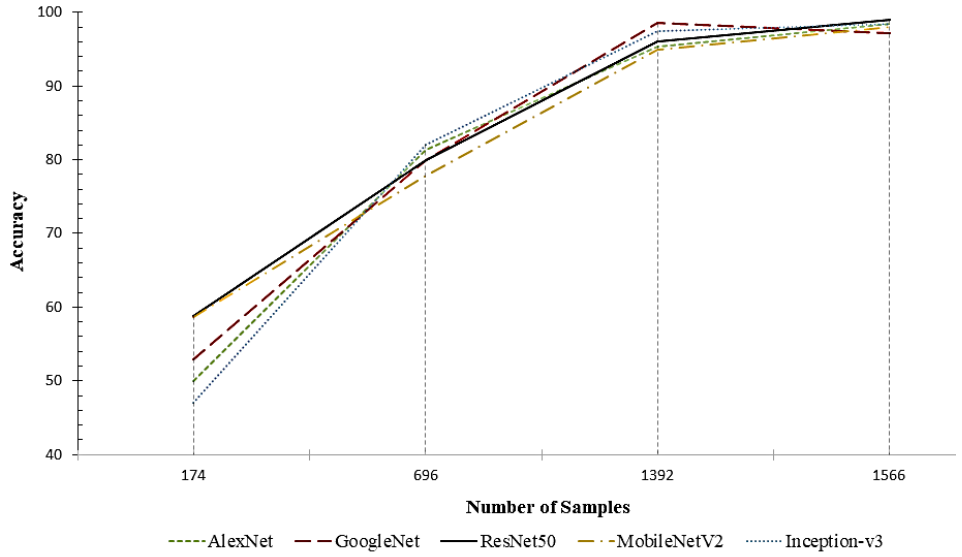

Fig 8. Influence of number of samples on the performance of the TL models for experiments 1, 2, 3 and 4. ResNet50 achieved the highest accuracy with %99.02 in experiment 4.

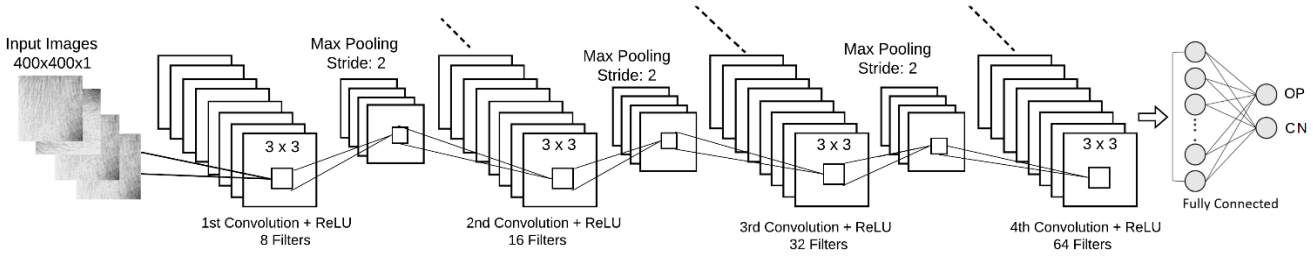

Fig 9. The CNN model used to classify dataset C. CL and FCL stands for Convolutional Layer and Fully Connected Layer, respectively

3.4. Experiment 4: Classification of Dataset C

In the previous experiments, the datasets did not include the original images. Here, the N original images were added to dataset B, providing a total of 1566 images. Tables 3 and 4 regroup the achieved results. Accuracy values continued to increase, except for GoogleNet, which was the only model with a less accuracy value compared to experience 3 (dataset B). Meanwhile, ResNet50 achieved an accuracy of 99.02%, followed by AlexNet and InceptionV3 with an accuracy of 98.4%. As shown in Table 4, AlexNet correctly classified all the CN samples (FN = 0), while ResNet50 and InceptionV3 correctly classified all the OP samples (FP = 0). As an overall statement, ResNet50 and InceptionV3 achieved the best classification scores, followed by GoogleNet that showed an impressive performance at the third experiment. Figure 8 illustrates a comparison between the different TL models performances. It also shows the influence of data augmentation on the classification task.

3.5. Experiment 5: Training a simple CNN model from Scratch

In addition to classifying C dataset (1566 images) with AlexNet, MobileNetV2, GoogleNet, InceptionV3, and Resnet50, we also implemented a simple CNN model to see how it would perform. The model consisted of 5 layers; 4 convolutional layers and one fully connected layer. The first 3 convolutional layers were followed by a max pooling layer, and the fully connected layer was followed by a soft max function to normalize the fully connected layer's output. The proposed CNN can be seen in figure 9. The training options in table 2 were held the same in this experiment as well. The obtained results are shown in table 5. As can be seen, with an Acc of 96.1%, results were surprisingly close to these obtained using Transfer Learning models, this demonstrates the efficiency of the implemented data augmentation procedure.

Table 5. Obtained Acc, Sp and Sn rates using a simple CNN model and dataset C.

Acc	Sp	Sn
96.1%	95.5%	96.8%

3.6. Experiment 6: Max (Majority) Voting Approach

Max-Voting is an ensemble learning method that is used to minimize false classified data samples and eventually increase the accuracy of the model. In machine learning, ensemble learning methods are used to combine multiple classifiers to make a decision [39-40]. Here, decisions from pre-trained models and the 5-layer CNN model from the previous experiment were combined and only samples that were classified correctly by the majority of the models (in this case the majority is 4) are considered as correctly classified. Samples that were misclassified by more than 2 models were not considered. In this experiment, dataset C was used. The number of test samples was 157 images per class (314 in total). Using this approach, an accuracy of 99.02% was achieved, which is the same results achieved by ResNet50 in Experiment 4. The reason is that the three OP images that were misclassified by ResNet50 were also misclassified by the other 5 models (4 pre-trained models and one simple CNN model). Three of the OP images were voted as CN, while all the CN images were voted as CN. Figure 10 illustrates the majority voting method for making a final decision.

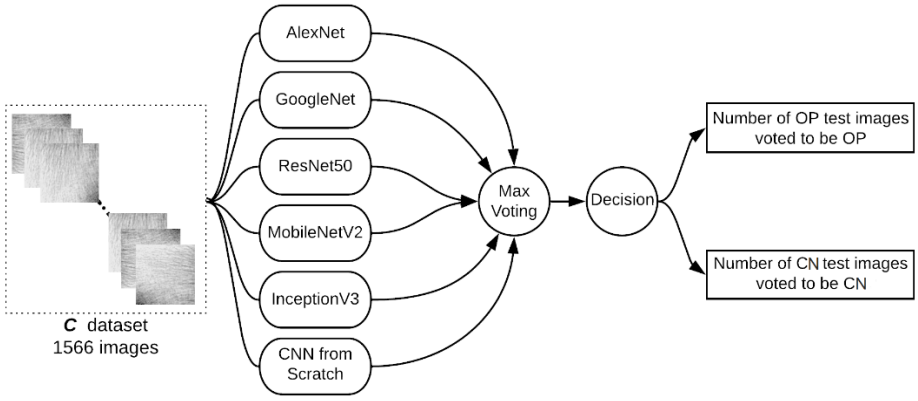

Fig 10. Max-Voting approach for making a final decision.

3.7. Experiment 7: 3-Fold Cross Validation

To validate the reached results and avoid overfitting issues, a K-fold Cross Validation (CV) approach was carried out as a final validation experiment. C dataset (1566 images) was subdivided into 3 subsets (see section 2.2.5) of 522 samples each. With 3-fold Cross Validation, each model was trained 3 times, and at the end of the training process, the mean of the obtained accuracies was computed. Table 6 presents the obtained mean accuracy for each model. As can be seen, GoogleNet managed to correctly classify all the test images at every training process, achieving an Acc of 100%. AlexNet followed with an Acc of 99.4%, which is the highest Acc that AlexNet achieved in all the experiments. Meanwhile, ResNet50, InceptionV3, MobileNetV2 and the simple CNN models performances diminished compared to those obtained using the same dataset in experiment 4 and 5.

3.8. Comparison to State-of-the-Art Studies

The proposed approach was also compared to some existing methods in the literature. Comparisons are reported in Table 7. A brief description of each method can be found in the Introduction. As can be seen, Transfer Learning was used in a different study where the accuracy value 84% was achieved. In this study, an accuracy of 99% was reached using ResNet50 thanks to data augmentation approach. Also, 100% accuracy was achieved with GoogleNet when the 3-fold cross validation was implemented.

Table 6. Obtained mean accuracies for each model with 3-fold cross validation and dataset C.

3-Fold CV	AlexNet	GoogleNet	ResNet50	MobileNetV2	InceptionV3	CNN from Scratch
1566 images	99.4%	100%	97.6%	91.8%	94.4%	91.7%

Table 7. Comparison to state-of-the-art studies for osteoporosis classification

Title of the study	Year	Data	Method	Acc Value
3D image analysis and artificial intelligence for bone disease classification [7]	2010	18 trabecular bone samples	Genetic Algorithm	100%
Texture Analysis and Genetic Algorithms for Osteoporosis Diagnosis [8]	2010	18 trabecular bone samples	Genetic Algorithm	100%
One Dimensional Local Binary Pattern for Bone Texture Characterization [9]	2011	80 X-Ray ROIs from calcaneus bone images	1D Local Binary Pattern 1DLBP	85% AUC
Early Diagnosis Of Osteoporosis Using Artificial Neural Networks And Support Vector Machines [10]	2012	80 computed tomography (CT) images	Support Vector Machine	86%
Osteoporosis Assessment Using Multilayer Perceptron Neural Networks [11]	2012	120 X-Ray ROIs calcaneus bone images	Multilayer Perceptron	97%
Evaluation of fractional Brownian motion synthesis methods using the SVM classifier [13]	2014	77 X-Ray ROIs from calcaneus bone images	Support Vector Machine	95%
Diagnosis of osteoporosis disease from bone X-ray images with Stacked Sparse Auto-encoder and SVM classifier [14]	2017	174 X-Ray ROIs from calcaneus bone images	Support Vector Machine	95.5%
Towards Deep Learning Based Estimation of Fracture Risk in Osteoporosis Patients [15]	2017	Synthetic data 25000 trabecular bone	Convolutional Neural Network	93.8%
Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans [16]	2018	1432 CT scans	CNN + Long Short Term Memory	89.2%
Osteoporosis identification based on the validated trabecular area on digital dental radiographic images [17]	2019	84 ROIs from 42 digital dental radiographic images	Multilayer Perceptron	87.87%
Integrative Blockwise Sparse Analysis for Tissue Characterization and Classification [18]	2020	174 X-Ray ROIs from calcaneus bone images	maximum a posteriori probability (BBMAP) and log likelihood function (BBLL)	100%

Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses [19]	2020	174 X-Ray ROIs from calcaneus bone images	Fusing CNN features with hand-crafted features	77.5%
Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs [20]	2020	680 patients	TL - VGG16	84%
Detection and Segmentation of Osteoporosis in Human Body using Recurrent Neural Network [41]	2020	Hand X-rays	LeNet based CNN	93.9%
Detection of Osteoporosis in Defected Bones Using Radtorch and Deep Learning Techniques [42]	2021	1000 bone X-rays	TL – ResNet50	98%
Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography [43]	2021	Pelvic X-rays	TL - ResNet18	88%
Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning [44]	2021	Pelvis and Lumbar spine X-rays	TL - VGG16	91.7% for hip osteoporosis 86.2% for spine osteoporosis
This Study	2021	Augmented data 1566 X-Ray ROIs from calcaneus bone images	TL - GoogleNet + 3-Fold Cross Validation	100%
This Study	2021	Augmented 1566 X-Ray ROIs from calcaneus bone images	TL - ResNet50 - 80% training, 20% test	99.02%

4. Conclusion

Due to the high similarity between healthy and osteoporotic patients bone microarchitecture in X-ray images, osteoporosis classification using deep learning models is a very challenging task. Despite that, using an adequate data augmentation strategy and deep transfer learning architectures can be a new opportunity to improve the diagnosis of osteoporosis. Achieved outstanding results indicate that the applied contrast, brightness, and sharpness manipulation provided the retained models with discriminatory information that is important for the classification of the two populations of healthy and osteoporotic patients.

First, both 90 and 270 degrees rotation were implemented on the original data to provide more training data and increase the generalizability of the model. Contrast, brightness, and sharpness adjustments reduced the similarity between control and osteoporotic images and brought to light more discriminatory information by disclosing the details of the bone microarchitecture. The impact of data augmentation can intelligibly be seen throughout the classification experiments in section 3. Firstly, the set of original data (174 images) was classified, and accuracy values varied between 47% and 58.8% using the retained learning models. Secondly, using the dataset A with 696 images, accuracy values were increased to vary from 77.8% to 82.1%. Thirdly, the dataset B with 1392 images was classified and the model's accuracy was boosted to 94.9% with MobileNetV2 and 98.5% with GoogleNet. The best classification results were obtained using the dataset C, which consisted of 1566 images. In this case an accuracy of 99.02% was accomplished with the ResNet50 model.

Examining the obtained results using ResNet50, the accuracy of 58.8% reached using the original dataset (174 images) was improved to 80% with dataset A (696 images). Then to 96% using dataset B (1392 images). To finally reach an accuracy of 99.02% with dataset C (1566 images). This classification accuracy improvement demonstrates the positive influence of the proposed data augmentation approach on the performance of the models. Furthermore, all the test images were correctly classified when a 3-fold cross-validation was used, leading to an accuracy of 100% using GoogleNet model.

To put the used data augmentation procedure in a more serious test, dataset C was classified with a simple CNN model trained from scratch. The model accomplished a promising Acc of 96.1%, showing the contribution of the proposed data augmentation approach. Examining similar works in the literature shows that leading results in the task of osteoporosis classification using CNNs [13, 14, 17] or Transfer Learning [18] are reached in this study (100% with GoogleNet). As a conclusion, we believe that an approach with such data augmentation methodology and model training options can make a remarkable contribution to the early diagnosis of osteoporosis.

References

[1] Tuck SP, Francis RM: Osteoporosis. Postgrad Med J 78(923), 526-532, 2002 Apr

[2] Christodoulou C, Cooper C: What is osteoporosis?. Postgraduate medical journal 79(929), 133-8, 2003 Mar

[3] Cooper C, Campion G, Melton L3: Hip fractures in the elderly: a world-wide projection. Osteoporosis International 1;2(6):285-9, 1992 Nov

[4] Sözen T, Özışık L, Başaran NÇ: An overview and management of osteoporosis. European journal of rheumatology 4(1) 46, 2017 Mar

[5] National Osteoporosis Foundation. Available at https://www.nof.org/patients/diagnosis-information/bone-density-examtesting. Accessed 14 June 2020

[6] Jennane R, Harba R, Lemineur G, Bretteil S, Estrade A, Benhamou CL: Estimation of the 3D self-similarity parameter of trabecular bone from its 2D projection. Medical Image Analysis 11.1:91-98, 2007 Feb

[7] Akgundogdu A, Jennane R, Aufort G, Benhamou CL, Ucan ON: 3D image analysis and artificial intelligence for bone disease classification. Journal of medical systems 34.5:815-828, 2010 Oct

[8] Yousfi L, Houam L, Boukrouche A, Lespessailles E, Ros F, Jennane R: Texture Analysis and Genetic Algorithms for Osteoporosis Diagnosis. International Journal of Pattern Recognition and Artificial Intelligence 6;34(05):2057002, 2020 May

[9] Houam L, Hafiane A, Boukrouche A, Lespessailles E, Jennane R: One Dimensional Local Binary Pattern for Bone Texture Characterization. Pattern Analysis and Applications, Springer, Volume 17, Issue 1, pp. 179-193, 2014

[10] İstanbullu M, Aydin M, Benveniste R, Uçan ON, Jennane R: Early diagnosis of osteoporosis using artificial neural networks and support vector machines, 2012 20th Signal Processing and Communications Applications Conference (SIU). IEEE, 2012

[11] Harrar K, Hamami L, Akkoul S, Lespessailles E, Jennane R: Osteoporosis assessment using Multilayer Perceptron neural

networks, 2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2012

[12] Touvier J, Winzenrieth R, Johansson H, Roux JP, Chaintreuil J, Toumi H, Jennane R, Hans D, Lespessailles E: Fracture discrimination by combined bone mineral density (BMD) and microarchitectural texture analysis. Calcified tissue international 96.4:274-83, 2015 Apr

[13] Tafraouti A, El Hassouni M, Jennane R: Evaluation of fractional Brownian motion synthesis methods using the SVM classifier. Biomedical Signal Processing and Control 1;49:48-56, 2019 Mar

[14] Nasser Y, El Hassouni M, Brahim A, Toumi H, Lespessailles E, Jennane R: Diagnosis of osteoporosis disease from bone X-ray images with stacked sparse autoencoder and SVM classifier, 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2017

[15] Ciuşdel CF, Vizitiu A, Moldoveanu F, Suciu C, Itu LM: Towards deep learning based estimation of fracture risk in osteoporosis patients, 2017 40th International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2017

[16] Tomita N, Cheung YY, Hassanpour S: Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Computers in biology and medicine 98:8-15, 2018 Jul

[17] Sela EI, Pulungan R: Osteoporosis identification based on the validated trabecular area on digital dental radiographic images. Procedia Computer Science 157:282- 289, 2019 Jan

[18] Zheng K, Harris CE, Jennane R, Makrogiannis S: Integrative Blockwise Sparse Analysis for Tissue Characterization and Classification. Artificial Intelligence in Medicine 1:101885, 2020 Jun

[19] Su R, Liu T, Sun C, Jin Q, Jennane R, Wei L: Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 14;385:300-9, 2020 Apr

[20] Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J: Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs. Journal of Clinical Medicine 9.2:392, 2020 Feb

[21] [dataset] Lespessailles E, Gadois C, Kousignian I, Neveu J.P, Fardellone P, Kolta S, Roux C, Do-Huu J.P, Benhamou C.L: Clinical interest of bone texture analysis in osteoporosis: a case control multicenter study. Osteoporosis International 19 1019–1028, 2008

[22] LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521(7553):436-44, 2015 May

[23] Perez L, Wang J: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv 1712.04621, 2017 Dec 13.

[24] Cao G, Huang L, Tian H, Huang X, Wang Y, Zhi R: Contrast enhancement of brightness-distorted images by improved adaptive gamma correction. Computers & Electrical Engineering 66: 569-582, 2018 Feb

[25] Zhang B, Allebach JP: Adaptive bilateral filter for sharpness enhancement and noise removal. IEEE transactions on Image Processing 31;17(5):664-78, 2008 Mar

[26] Rawat W, Wang Z: Deep convolutional neural networks for image classification, A comprehensive review. Neural computation.;29.9: 2352-2449, 2017 Sep

[27] Nwankpa C, Ijomah W, Gachagan A, Marshall S: Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv 1811.03378, 2018 Nov

[28] Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv 1409.1556, 2014 Sep

[29] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016

[30] He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770-778, 2016

[31] Krizhevsky A, Sutskever I, Hinton GE: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems pp. 1097-1105, 2012

[32] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 1-9, 2015

[33] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC: Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4510-4520, 2018

[34] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ: Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4700-4708, 2017

[35] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L: Imagenet A large-scale hierarchical image database. Computer Vision and Pattern Recognition, CVPR 2009. IEEE Conference on pp. 248-255, 2009 Jun

[36] Ruder S: An overview of gradient descent optimization algorithms. arXiv preprint arXiv 1609.04747, 2016 Sep

[37] Distill. Available at https://distill.pub/2017/momentum. Accessed 09 May 2020

[38] Qian N: On the momentum term in gradient descent learning algorithms. Neural networks 12.1: 145-151, 1999 Jan

[39] Sagi O, Rokach L: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1249, 2018

[40] Opitz D, Maclin R: Popular ensemble methods: An empirical study. Journal of artificial intelligence research 1;11:169-98, 1999

[41] Atheel Sabih Shaker, "Detection and Segmentation of Osteoporosis in Human Body using Recurrent Neural Network", International Journal of Advanced Science and Technology 2020;29(02):1055 1066.

[42] Shahzad M, Khan TF, Bashir M, Ayub M, Ashraf F, Hashmi S, Zahoor F, Jaskani FH. DETECTION OF OSTEOPOROSIS IN DEFECTED BONES USING RADTORCH AND DEEP LEARNING TECHNIQUES. International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 115-123

[43] Ho CS, Chen YP, Fan TY, Kuo CF, Yen TY, Liu YC, Pei YC. Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography. Archives of Osteoporosis. 2021 Dec;16(1):1-2.

[44] Hsieh CI, Zheng K, Lin C, Mei L, Lu L, Li W, Chen FP, Wang Y, Zhou X, Wang F, Xie G. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. Nature communications. 2021 Sep 16;12(1):1-9.