Deep Learning for Osteoporosis Classification Using Convolutional Neural Network on Knee Radiograph Dataset to Compare Classification Accuracy between RGB Images and Grayscale Images

Usman Bello Abubakar[1], Moussa Mahamat Boukar[2], Steve Adeshina[3], Senol Dane[4*]

[1]Department of Computer Science, Baze University, Abuja, Nigeria
[2]Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria
[3]Department of Computer Engineering, Nile University of Nigeria, Abuja, Nigeria
[4]Department of Physiology, Nile University of Nigeria, Abuja, Nigeria

ABSTRACT

Osteoporosis is a silent killer disease in the elderly that causes bone fragility and fractures. An early and precise diagnosis of osteoporosis is necessary to save the patient's life. Osteoporosis, along with bone fracture: its clinical manifestation is a complex illness towards which a great deal of research has been made upon. Machine Learning (ML) and Deep Learning (DL) advancements have allowed the area of Artificial Intelligence (AI) to make significant progress in complicated data environments where humans' capacity to find high-dimensional connections is limited. This work focuses on determining the appropriate image format to be used by a deep learning model to predict osteoporosis from knee radiograph. The image format considered in the study is RGB images and Grayscale images. Two Convolutional Neural Network (convent) of the same structure were trained on set of same images with RGB and grayscale. The result showed that the convent trained on grayscale images had a better accuracy than the convent trained on RGB images.

Keywords: Osteoporosis, Fracture risk assessment tool, Dual-energy X-ray absorptiometry, Bone mineral density

INTRODUCTION

Osteoporosis is a complex disease in which bone quantity and quality decline, increasing bone fragility". The clinical result of osteoporosis fracture affects 8.9 million people globally every year, posing a significant health, societal, and economic burden [1]. The bone, which gives shape to the entire body structure, is made up of tissues composed of organic and inorganic elements. Bone remodeling is the process of maintaining a healthy skeleton by the replacement of old bone with new bone. Bone loss occurs when a smaller volume of bone is replaced for a bigger volume of bone removed. This loss of bone tissue causes skeletal architecture to become disorganized, increasing the risk of fracture by

making the bone fragile. This imbalance that causes such bone fragility is primarily common in people above the age of 80 or simply menopausal women [2-4].

The measurement of Bone Mineral Density (BMD) by dual-energy X-ray absorptiometry is used to detect osteoporosis (DXA). BMD testing is the most dependable method of screening for osteoporosis. Bone strength is calculated by multiplying "Bone Mineral Content (BMC) by bone area". It is the quantity of bone mass per unit volume or area as shown in equation [5].

$$BMD\ (g/cm^2) = (BMC\ g\ /\ area(cm^2))$$

The examinations are generally performed on skeletal locations, such as the lumbar spine and the femoral neck [6]. The most common procedure for confirming an osteoporosis diagnosis is DXA. The World Health Organization (WHO) recommends that BMD measures be taken primarily at the hip and spine. Based on BMD measurements obtained by DXA, T-scores and Z-scores as reported by BMD test results have been defined by WHO and National Osteoporosis Foundation (NOF) as osteoporosis criteria [7].

However, "this technique does not take into account clinical risk factors or other bone metrics such as trabecular bone score, geometry, and so on. Old age, gender: female, ethnicity, inheritance, previous fracture, malnutrition, alcohol intake, current smoking, vitamin D deficiency, physical inactivity, various medicines, and medical disorders are the main clinical risk factors for osteoporotic fractures and altered bone metabolism." The Fracture Risk Assessment Tool (FRAT) is the most widely used and well-established clinical approach for fracture prediction (FRAX). It is founded on a number of previously established clinical risk factors, as well as BMD [8]. As indicated in Figure 1, osteoporosis is a complicated disorder that can be linked to a number of factors, including demographic information, patient clinical records pertaining to disease diagnoses and treatments, family history, diet, and lifestyle [9].

Artificial Intelligence (AI) has been employed as a screening tool and as an adjunct technology for image classification [10]. According to a review study published in 2019, recent advances in AI have resulted in practical applications that aid in the diagnosis of osteoporosis [11]. Osteoporosis has been classified with AI using dental radiographs [12, 13] spine radiographs [14], hand and wrist radiographs [15, 16].

Literature Review

Wrapper-based feature selection was used to compare classification algorithms for osteoporosis prediction. Multilayer Feed-forward Neural Networks (MFNN) [17], Nave Bayes, and logistic regression were employed as classification techniques. To find a subset of important SNPs, a wrapper-based feature selection method was applied. The MFNN model using the wrapper-based method was found to be the best predictive model for inferring disease risk in Taiwanese women based on the complex relationship between osteoporosis and SNPs. Patients "and doctors can use the proposed technology to improve decision-making based on clinical factors such as SNP genotyping data, according to the findings."

The study proposed by [18] "investigates whether using clinical data in conjunction with deep learning to diagnose osteoporosis from hip radiographs improves

Figure 1: Osteoporosis Risk Factors.

diagnostic performance over using the image mode alone." Between 2014 and 2019, they obtained 1131 images for objective labeling from individuals who had skeletal bone mineral density testing as well as hip radiography at the same medical hospital. Five Convolutional Neural Network (CNN) models were used to assess osteoporosis from hip radiographs. They also looked into ensemble models, which contained clinical characteristics in each CNN. The following were assigned to each network

Accuracy

Precision

Recall

Specificity

Negative Predictive Value (NPV)

F1 score, and

Area under Curve.

The five CNN models were tested using only hip radiographs, and Google Net and Efficient Net b3 had the highest accuracy, precision, and specificity. When patient variables were incorporated, Efficient Net b3 had the highest accuracy, recall, NPV, F1 score, and AUC score among the five ensemble models. When clinical factors from patient records were included to the CNN models, they improved even more in detecting osteoporosis from hip radiographs.

Deep learning techniques for screening osteoporosis with DPR pictures were created and assessed by Lee ki sun et al [19]. This study tested various CNN models based on osteoporosis discriminating accuracy using classified panoramic radiograph images based on the BMD value (T-score). The impact of transfer learning and fine-tuning a deep CNN model on classification performance was also investigated. Deep CNNs have been discovered to be effective image classification tools, but they require a huge quantity of training data, which may make them challenging to apply to medical radiographic image data. Transfer learning is a powerful approach for training deep CNNs without overfitting when the target dataset is significantly smaller than the basis dataset. The most common method for transfer learning is to use pertained models in a two-step process: On a large general dataset, the first n layers of a pretrained base network are copied to the first n levels of a target network, and then the remaining layers of the target network are randomly begun and trained toward the target task on a small local dataset [20]. "Several state-of-the-art results based on transfer learning algorithms showed outperformance in both general image classification and object recognition [21]."
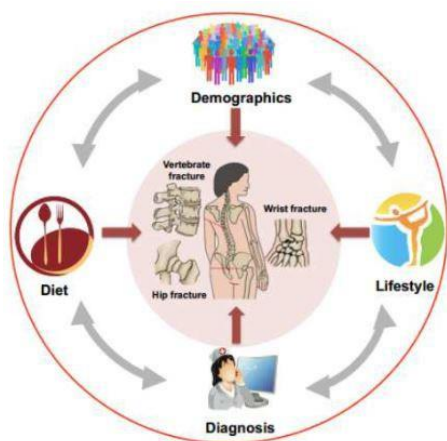
MATERIALS AND METHODS

Data Acquisition

he dataset comprises of 57 normal knee radiograph

images and 323 osteoporotic knee radiograph images of patients acquired from Kaggle: an online deep learning data repository. Table 1 below shows the splitting of image data into train-test-val folder. This project structuring is recommended so as to properly train the Deep Learning model.

Data Augmentation

Overfitting can be reduced by using a technique called data augmentation. When a machine learning model or neural network learns a function with a very large variance in order to perfectly model the training data, this is known as overfitting.

Image data augmentation is a technique for artificially increasing the size of a training dataset by making modified versions of the images in the dataset. Deep learning neural network models that have been trained on more data are more accurate, and augmentation approaches can provide image variations that boost the models' ability to generalize what they've learned to new images [22].

For this study, the Keras Image Data Generator python class was used to perform data augmentation. It allows for quickly and easily enhancing images using a variety of augmentation techniques as itemized below

Standardization

Rotation

Shifts

Flips

Brightness changes, among others.

The key advantage of utilizing the Keras Image Data Generator class, however, is that it is intended to give real-time data augmentation. At each epoch, the Image Data Generator class guarantees that the model receives new variations of the images. The Image Data Generator implementation for our data augmentation is shown in Figure 2.

Image Scaling

The pixel values in most image data are integers with values ranging from 0 to 255. Inputs with big integer values can disrupt or slow down the learning process in neural networks, which process inputs with modest weight values. As a result, picture normalization is a recommended practice: normalize image pixel values so that each pixel value is between 0 and 1.

Image pixel values are normalized by dividing all pixel values by the biggest pixel value, which are 255. This is done across all channels, regardless of the image's actual range of pixel values. The image was normalized for numerical stability and convergence. This ensures the neural network has a higher chance of converging and the gradient descent algorithm is way more likely to be stable.

The images in the dataset were normalized (rescaled) using the Image Data Generator method and passing rescale=1. /255 as its argument.

Image Resizing

In computer vision, resizing images is an important preprocessing step. Machine learning algorithms, in general, train quicker on smaller images. When an input image is twice as large, a network must learn from four times as many pixels, which takes time. Furthermore, the designs of deep learning models demand that our images be the same size, and most raw collected images vary in size.

The size of the images in our dataset is relatively large and varies in size. This sort of inconsistency would result in poor accuracy of our network and thus, the images all need to be of the same height and width. The ideal way to deal with different sized images is to downscale them to match dimensions from the smallest image available.
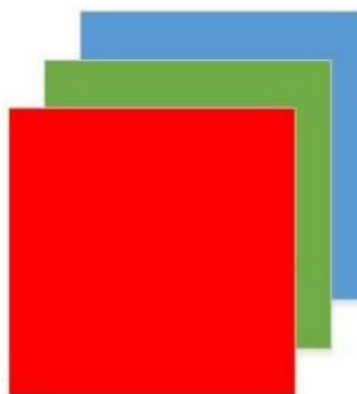
```
##########################################
train = ImageDataGenerator(
    rescale=1/255,
    horizontal_flip=True,
    vertical_flip=True,
    zoom_range=0.2, # zoom
    rotation_range=10,
    fill_mode='nearest'
)
```

Figure 2: Data Augmentation.

Figure 3: RGB Image Representation.

Table 1: Image Distribution.

Category	Total (T)	Training (0.8*0.8*T)	Validation (0.2*0.8*T)	Testing (0.2*T)
0(Normal)	57	36	9	11
1(osteoporosis)	323	207	52	65

Image Data Format

The two image formats considered in this study are RGB images and Grayscale images. The dataset consists of images in RGB format. An RGB (red, green, blue) image is a three-dimensional byte array in which each pixel's color value is explicitly stored. RGB image arrays are made up of three-color channels and a width and height [23]. As demonstrated in Figure 3, an RGB image can be regarded logically as three separate images (a red scale image, a green scale image, and a blue scale image) placed on top of each other.

An image in RGB format increases the complexity of the model. This is the reason why gray scale images are preferred over colored ones to simplify mathematics. It is relatively easier to deal with (in terms of mathematics) a single-color channel (shades of white/black) than multiple color channels (shades of Red, Green, and Blue).

A monochrome image, often known as a one-color image, is a grayscale image. As illustrated in Figure 4, it simply provides brightness information and no color information. The grayscale data matrix values are then used to indicate intensities. A common image has 8 bits per pixel, allowing it to represent (0-255) distinct brightness (gray) levels [24]. And Figure 5 shows the code that was used to iteratively convert images to grayscale.

From Figure 5, the variables source path and destination path serves as the path to RGB images and path to store new grayscale images respectively. The code in the image depicts the conversion of images in the train folder for osteoporosis cases. The same code was executed for the images in the test and Val folder as appropriate. Figure 6 shows an RGB image on the left and its corresponding grayscale equivalent on the right.

CNN Model Architecture

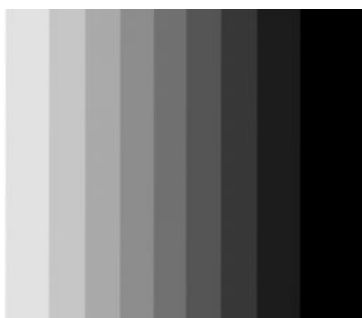The most widely used deep learning architecture is

Figure 4: Grayscale Image Representation.

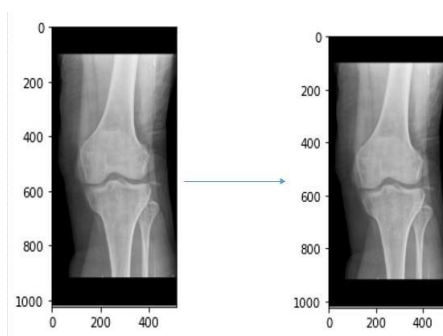**Figure 5: RGB to Grayscale Python Code.**

Figure 6: RGB and Grayscale Equivalent.

Convolutional Neural Networks (CNNs). CNN is mostly utilized in image-related tasks. CNN's popularity has grown as a result of its capacity to extract essential patterns from photos without the involvement of humans. For example, a trained CNN model with enough instances of photos may automatically determine whether a given image is a dog or cat for a given data set of dog and cat. This is achievable by extracting some important features in different layers from the pixel information of the image. CNN model is efficient in image classification, and it has composed layers of convolution, pooling and fully connected layers.

Two proposed CNN architecture were implemented: one for RGB image and another for Grayscale images. However, the architecture of the CNN was similar with the exception of the input layer due to the difference in number of channels in RGB as opposed to grayscale. Three convolutional layers, two max-pooling layers, and a fully connected layer constituted the network. The images of input shape 1000 x 1000 were sent into the network as input. The conditional probability distribution p (YX) over the two categories of osteoporosis (1) and normal was the network's output (0). The network is made up of three convolutional layers, each followed by a max pooling layer and finally a ReLU activation layer. In the grayscale convolutional layer, the kernels were 5x5 and in the RGB convolutional layer, they were 5 x 5 x 3. Two fully connected layers with a sigmoid function followed the final convolutional layer.

K-Fold Cross Validation

Cross-validation is a technique used to prevent overfitting in a prediction model, which is especially useful when the amount of data available is restricted. In cross-validation, a value for K is chosen, and the dataset is divided into K equal-sized partitions. After splitting, K–1 partitions are used for training and one for testing. This method is done K times, each time with a distinct partition utilized for testing. This study makes use of 5-fold cross validation.

RESULTS

Model Evaluation on RGB Images

The CNN was tested on unseen data using the Keras evaluate function. The function was passed one argument

which served as the path to the folder containing the testing images. The function returned a test loss of 52% and a test accuracy of 83% as seen in Figure 7. During model training, for evaluating the model on the RGB image, a graph was created to denote a plot of accuracy on the training and validation datasets over training epochs as shown in Figure 8.

Model Evaluation on Grayscale Images

The CNN was tested on unseen data using the Keras evaluate function. The function was passed one argument which served as the path to the folder containing the testing images. The function returned a test loss of 60% and a test accuracy of 86% as seen in Figure 9. During model training, for evaluating the model on the grayscale image, a chart was created to denote a plot of accuracy on the training and validation datasets over training epochs as shown in Figure 10.

Confusion Matrix

A confusion matrix is a table that summarizes the results of classification problem prediction. Count values are used to sum and break down the number of correct and incorrect predictions by class. The confusion matrix shows the different ways in which the network model gets confused while making predictions. The confusion matrix for the convents generated by the python scikit learn package, as denoted in the snippet given in Figure 11, is presented in Figure 12 and 13.


```
test_loss, test_acc = model.evaluate(testGen)
print('Test loss: {} Test Acc: {}'.format(test_loss, test_acc))

[==============================] - 45s 3s/step - loss: 0.5194 - accuracy: 0.8313
Test loss: 0.519477336614575 Test Acc: 0.831361735801697
```

Figure 7: Model test accuracy and test loss on RGB Images.

Figure 8: Graph of model accuracy on RGB Images.

```
test_loss, test_acc = model.evaluate(testGen)
print('Test loss: {} Test Acc: {}'.format(test_loss, test_acc))

[==============================] - 45s 3s/step - loss: 0.5903 - accuracy: 0.8606
Test loss: 0.5903562346614575 Test Acc: 0.860661724763597
```

Figure 9: Model test accuracy and test loss on Grayscale Images.

Figure 10: Model accuracy on Grayscale Images.

```
from sklearn.metrics import classification_report, confusion_matrix
import seaborn as sns
Y_pred = model.predict_generator(testGen, steps=(76 // my_batch_size+1))
y_pred = np.argmax(Y_pred, axis=1)
print('Confusion Matrix\n')
print(y_pred)
#print(testGen.classes)
cm = confusion_matrix(testGen.classes, y_pred)
plt.figure(figsize=(10,7))
sns.heatmap(cm,annot=True)
plt.xlabel("Predicted")
plt.ylabel("Truth")
plt.show()
```
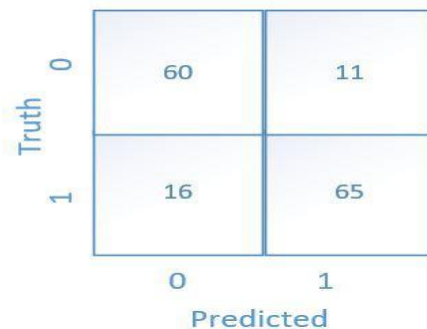
Figure 11: Confusion Matrix code.

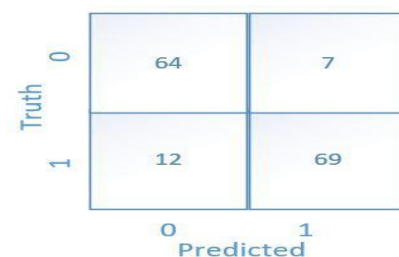
Figure 12: Confusion Matrix for RGB Convent.

Figure 13: Confusion Matrix for Grayscale Convent.

Scikit Learn Classification Report on Grayscale convent

Precision

Precision is the ratio between the True Positives and all the Positives in Figure 14. It "is the measure of patients that we correctly identify having osteoporosis disease out of all the patients actually having it."

Recall

The "recall is the measure of our model correctly

Figure 14: Classification Report Table.

identifying True Positives. Thus, for all the patients who actually have osteoporosis disease, recall tells us how many we correctly identified as having an osteoporosis disease."

F1-Score

This "is a combination of precision and recall. It is the harmonic mean of the model's precision and recall."

CONCLUSION

This study proved that CNNs can detect osteoporosis from knee radiographs with a reasonable degree of accuracy. The convent that was applied on grayscale images performed slightly better than the convent that was applied on RGB images. The accuracy of the convent for grayscale images was 86% while that of RGB images was 83%.

The slight increase is due to the fact that, grayscale image has a smaller number of channels than RGB and thus, RGB increases the complexity of a network model. Also, the slight increase would be that despite the RGB images are composed of 3 channels, knee radiograph images are generally monochrome and thus, the RGB channels are not too pronounced. The difference in the accuracy is not much but in Deep Learning for medical diagnosis, any slight improvement in model accuracy is significant. The convents were trained utilizing the GPU on google colabs.

REFERENCES

1. Moreno MA, Skoracki RJ, Hanna EY, et al. Microvascular free flap reconstruction versus palatal obturation for maxillectomy defects. Head Neck 2010; 32:860-8.

2. Ali MM, Khalifa N, Alhajj MN. Quality of life and problems associated with obturators of patients with maxillectomies. Head Face Med 2018; 14:1-9.

3. Newton JT, Fiske J, Foote O, et al. Preliminary study of the impact of loss of part of the face and its prosthetic restoration. J Prosthet Dent 1999; 82:585-90.

4. Olson ML, Shedd DP. Disability and rehabilitation in head and neck cancer patients after treatment. Head Neck Surg 1978; 1:52-8.

5. Lethaus B, Lie N, De Beer F, et al. Surgical and prosthetic reconsiderations in patients with maxillectomy. J Oral Rehabil 2010; 37:138-42.

6. Davison SP, Sherris DA, Meland NB. An algorithm for maxillectomy defect reconstruction. Laryngoscope 1998; 108:215-9.

7. Hassona Y, Taimeh D, Marahleh A, et al. YouTube as a source of information on mouth (oral) cancer. Oral Dis 2016; 22:202-8.

8. Delli K, Livas C, Vissink A, et al. Is YouTube useful as a source of information for Sjögren's syndrome?. Oral Dis 2016; 22:196-201.

9. Şallı GA, Egil E. Are YouTube videos useful as a source of information for oral care of leukemia patients. Quintessence Int 2020; 51:78-85.

10. Desai T, Shariff A, Dhingra V, et al. Is content really king? An objective analysis of the public's response to medical videos on YouTube. PLOS 2013; 8:e82469.

11. Kumar N, Pandey A, Venkatraman A, et al. Are video sharing web sites a useful source of information on hypertension?. J Am Soc Hypertens 2014; 8:481-90.

12. Nason K, Donnelly A, Duncan HF. YouTube as a patient-information source for root canal treatment. Int Endod J 2016; 49:1194-200.

13. Adhikari J, Sharma P, Arjyal L, et al. YouTube as a source of information on cervical cancer. N Am J Med Sci 2016; 8:183.

14. Sorensen JA, Pusz MD, Brietzke SE. YouTube as an information source for pediatric adenotonsillectomy and ear tube surgery. Int J Pediatr Otorhinolaryngol 2014; 78:65-70.

15. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis—a wakeup call?. Int J Rheumatol 2012; 39:899-903.

16. Dos Santos DM, de Caxias FP, et al. Oral rehabilitation of patients after maxillectomy. A systematic review. Br J Oral Maxillofac Surg 2018; 56:256-66.

17. Lethaus B, Lie N, De Beer F, et al. Surgical and prosthetic reconsiderations in patients with maxillectomy. J Oral Rehabil 2010; 37:138-42.

18. Yagci F. Evaluation of YouTube as an information source for denture care. J Prosthet Dent 2021.

19. Khan ML. Social media engagement: What motivates user participation and consumption on YouTube?. Comput Hum Behav 2017; 66:236-47.

20. Fortuna G, Schiavo JH, Aria M, et al. The usefulness of YouTube™ videos as a source of information on burning mouth syndrome. J Oral Rehabil 2019; 46:657-65.

21. Nason K, Donnelly A, Duncan HF. YouTube as a patient-information source for root canal treatment. Int Endod J 2016; 49:1194-200.

22. Mueller SM, Hongler VN, Jungo P, et al. Fiction, falsehoods, and few facts: cross-sectional study on the content-related quality of atopic eczema-related videos on YouTube. J Med Internet Res 2020; 22:e15599.

23. Ramadhani A, Zettira Z, Rachmawati YL, et al. Quality and reliability of halitosis videos on YouTube as a source of information. Dent J 2021; 9:120.

24. Coban G, Buyuk SK. YouTube as a source of information for craniofacial distraction osteogenesis. J Craniofac Sur 2021; 32: 2005-7.